8-2013

Finite Difference and Discontinuous Galerkin Finite Element Methods for Fully Nonlinear Second Order Partial Differential Equations

Thomas Lee Lewis
tlewis10@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Thomas Lee Lewis entitled "Finite Difference and Discontinuous Galerkin Finite Element Methods for Fully Nonlinear Second Order Partial Differential Equations." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Mathematics.

Xiaobing H. Feng, Major Professor

We have read this dissertation and recommend its acceptance:

Suzanne M. Lenhart, Ohannes Karakashian, Alfredo Galindo-Uribarri

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

8-2013

Finite Difference and Discontinuous Galerkin Finite Element Methods for Fully Nonlinear Second Order Partial Differential Equations

Thomas Lee Lewis
tlewis10@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Thomas Lee Lewis entitled "Finite Difference and Discontinuous Galerkin Finite Element Methods for Fully Nonlinear Second Order Partial Differential Equations." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Mathematics.

Xiaobing H. Feng, Major Professor

We have read this dissertation and recommend its acceptance:

Suzanne M. Lenhart, Ohannes Karakashian, Alfredo Galindo-Uribarri

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Finite Difference and Discontinuous Galerkin Finite Element Methods for Fully Nonlinear Second Order Partial Differential Equations

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Thomas Lee Lewis

August 2013

This dissertation is lovingly dedicated to my parents who have always encouraged and supported me.

Acknowledgements

I would like to thank all of the people that have contributed to my doctoral research. I first would like to express my sincerest gratitude to my advisor, Professor Xiaobing Feng, for his guidance and encouragement throughout my graduate studies. Dr. Feng involved me in research early in my graduate career, and he has always shared his knowledge and experiences so that I could better understand the rewards of pursuing an academic career. Dr. Feng's willingness to share his wisdom has both made this work possible and laid a foundation for a future in academics.

I would also like to thank the remaining members of my dissertation committee, Professor Ohannes Karakashian, Professor Susanne Lenhart, and Dr. Alfredo Galindo-Uribarri, for their willingness to serve on my committee and answer any questions that I had. Dr. Karakashian's leadership in the classroom contributed to much of my interest and background in numerical analysis. Dr. Lenhart's advice was indispensable when preparing research talks, and Dr. Galindo-Uribarri's encouragement at ORNL first planted the seed for an academic research career.

Finally, I would like to thank my friends and family for their patience and encouragement along the way. My friends were always available for a distraction and understanding during my absence. My parents were always able to help me forget the stresses associated with my doctoral research. My parents-in-law took me in, always offering encouragement and support. Most importantly, I cannot express the thankfulness I have for my wife, Beth, and the love she has provided me.

Abstract

The dissertation focuses on numerically approximating viscosity solutions to second order fully nonlinear partial differential equations (PDEs). The primary goals of the dissertation are to develop, analyze, and implement a finite difference (FD) framework, a local discontinuous Galerkin (LDG) framework, and an interior penalty discontinuous Galerkin (IPDG) framework for directly approximating viscosity solutions of fully nonlinear second order elliptic PDE problems with Dirichlet boundary conditions. The developed frameworks are also extended to fully nonlinear second order parabolic PDEs. All of the proposed direct methods are tested using Monge-Ampère problems and Hamilton-Jacobi-Bellman (HJB) problems. Due to the significance of HJB problems in relation to stochastic optimal control, an indirect methodology for approximating HJB problems that takes advantage of the inherent structure of HJB equations is also developed.

First, a FD framework is developed that guarantees convergence to viscosity solutions when certain properties concerning admissibility, stability, consistency, and monotonicity are satisfied. The key concepts introduced are numerical operators, numerical moments, and generalized monotonicity. One class of FD methods that fulfills the framework provides a direct realization of the vanishing moment method for approximating second order fully nonlinear PDEs. Next, the emphasis is on extending the FD framework using DG methodologies. In particular, some nonstandard LDG and IPDG methods that utilize key concepts from the FD framework are formulated. Benefits of the DG methodologies over the FD methodology include the ability to

handle more complicated domains, more freedom in the design of meshes, higher potential for adaptivity, and the ability to use high order elements as a means for increased accuracy. Last, a class of indirect methods for approximating HJB equations using the vanishing moment method paired with a splitting formulation of the HJB problem is developed and tested numerically. The proposed methodology is well-suited for both continuous and discontinuous Galerkin methods, and it complements the direct methods developed in the dissertation.

Table of Contents

Chapter 1

Introduction

1.1 Prelude

Partial differential equations (PDEs) provide a convenient mathematical language for describing relations between various quantities in a system. PDEs arise not only from other fields within mathematics such as differential geometry and analysis, but from almost every scientific and engineering field where mathematical models are used to describe some phenomena. In general, a second order PDE in spatial variables has the form

$$F\left(D^2u, \nabla u, u, x\right) = 0, \tag{1.1}$$

where $D^2u(x)$ and $\nabla u(x)$ denote the Hessian and gradient of u at x, respectively. PDEs are often classified based upon the nonlinearity of the PDE operator F. A fully nonlinear PDE corresponds to an equation where the operator F is nonlinear in the highest order derivative(s) appearing in the PDE. The theory for linear, semi-linear, and quasi-linear PDEs is well studied and can be considered classical in many situations. In contrast, fully nonlinear PDEs are still at the forefront of developing PDE analysis.

Closed form solutions do not exist for most PDEs, even for most linear PDEs. Thus, when a solution does exist for a PDE problem, in order to visualize the

solution(s), one must resort to numerical methods and algorithms to obtain an approximate solution with the help of a computer. Today, numerical PDEs has become a major research field in mathematics, largely driven by the vast array of applications. As with PDE theory, much of the numerical PDE theory for linear, semi-linear, and quasi-linear PDEs has been well developed and documented. However, due to the relative infancy of fully nonlinear PDE theory, the area of numerical PDEs for fully nonlinear PDEs is currently a growing area of interest.

The goal of this dissertation is to develop, analyze, and implement various numerical methods for directly and indirectly approximating viscosity solutions of fully nonlinear second order PDE problems with Dirichlet boundary conditions. We will focus on building a theoretical framework for designing direct finite difference (FD) and discontinuous Galerkin (DG) methods for approximating viscosity solutions of fully nonlinear second order PDEs. We will also develop an indirect methodology for approximating Hamilton-Jacobi-Bellman equations from stochastic optimal control. To provide proper context for our methodologies, we will recall results from PDE theory for fully nonlinear first and second order problems as well as the corresponding numerical PDE results for first order fully nonlinear problems.

1.2 PDE Solution Concepts and Facts

To prepare necessary background materials, we first present an overview of the relevant PDE theory for first and second order fully nonlinear PDEs. We begin with formally defining a PDE and the classification of PDE problems based upon degrees of nonlinearity.

Definition 1.1. *Fix an integer $k \geq 1$ and let Ω denote an open subset of $\mathbb{R}^d$. An expression of the form*

$$F\left(D^k u(x), D^{k-1}u(x), \ldots, Du(x), u(x), x\right) = 0, \qquad x \in \Omega \tag{1.2}$$

is called a k^{th}-order partial differential equation*, where*

$$F : \mathbb{R}^{d^k} \times \mathbb{R}^{d^{k-1}} \times \cdots \times \mathbb{R}^d \times \mathbb{R} \times \Omega \to \mathbb{R}$$

is given, and

$$u : \Omega \to \mathbb{R}$$

is the unknown.

Definition 1.2.

(i) The PDE (1.2) *is called* linear *if it has the form*

$$\sum_{|\alpha| \le k} a_\alpha(x) D^\alpha u = f(x)$$

for given functions f *and* a_α, $|\alpha| \le k$.

(ii) The PDE (1.2) *is called* semi-linear *if it has the form*

$$\sum_{1 \le |\alpha| \le k} a_\alpha(x) D^\alpha u + a_0\left(u, x\right) = 0$$

for given functions a_α, $|\alpha| \le k$.

(iii) The PDE (1.2) *is called* quasi-linear *if it has the form*

$$\sum_{|\alpha| = k} a_\alpha\left(D^{k-1}u, \ldots, Du, u, x\right) D^\alpha u + a_0\left(D^{k-1}u, \ldots, Du, u, x\right) = 0$$

for given functions a_0 *and* a_α, $|\alpha| = k$.

(iv) The PDE (1.2) *is called* fully nonlinear *if it depends nonlinearly upon the highest order derivatives.*

Since we are only concerned with first and second order PDEs, we let D^2 denote the Hessian operator and $\nabla := D^1$ denote the gradient operator in the following.

For presentation purposes, we adopt standard function and space notations as in [7] and [32]. For example, for a bounded open domain $\Omega \subset \mathbb{R}^d$, $B(\Omega)$, $USC(\Omega)$ and $LSC(\Omega)$ are used to denote, respectively, the spaces of bounded, upper semicontinuous, and lower semicontinuous functions on Ω. Also, for any $v \in B(\Omega)$, we define

$$v^*(x) := \limsup_{y \to x} v(y) \qquad \text{and} \qquad v_*(x) := \liminf_{y \to x} v(y).$$

Then, $v^* \in USC(\Omega)$ and $v_* \in LSC(\Omega)$, and they are called *the upper and lower semicontinuous envelopes* of v, respectively.

In presenting the relevant PDE theory for fully nonlinear first and second order PDEs, we will let H denote a general fully nonlinear first order operator and F denote a general fully nonlinear second order operator. More precisely, for $\Gamma \subsetneq \partial\Omega$, $H : \mathbb{R}^d \times \mathbb{R} \times (\Omega \cup \Gamma) \to \mathbb{R}$ and $F : \mathcal{S}^{d\times d} \times \mathbb{R}^d \times \mathbb{R} \times \overline{\Omega} \to \mathbb{R}$, where $\mathcal{S}^{d\times d}$ denotes the set of $d \times d$ symmetric real matrices. The general first and second order fully nonlinear PDE problems involve seeking locally bounded functions $u : (\Omega \cup \Gamma) \to \mathbb{R}$ and $v : \overline{\Omega} \to \mathbb{R}$ such that u and v are viscosity solutions (see Section 1.2.1) of

$$H(\nabla u, u, x) = 0 \qquad \text{in } \Omega \cup \Gamma, \tag{1.3a}$$

$$F(D^2 v, \nabla v, v, x) = 0 \qquad \text{in } \overline{\Omega}, \tag{1.3b}$$

respectively. We will also refer to (1.3a) as a Hamilton-Jacobi problem. For presentation purposes, we have used the convention of writing the boundary condition as a discontinuity of the PDE (cf. [3, p.274]). However, in our formulation of numerical methods for (1.3b), we will treat the boundary conditions explicitly when we assume Dirichlet boundary conditions.

A general solution theory that guarantees existence and uniqueness in some function class does not exist for fully nonlinear second order problems as represented by (1.3b) without some additional assumptions. Thus, we choose to impose some structure on fully nonlinear second order problems represented by (1.3b). More

precisely, we impose an ellipticity requirement. The following definition is standard (cf. [3, 7, 32]).

Definition 1.3. *Equation* (1.3b) *is said to be elliptic if, for all* $(\mathbf{p}, \lambda, x) \in \mathbb{R}^d \times \mathbb{R} \times \overline{\Omega}$*, there holds*

$$F(A, \mathbf{p}, \lambda, x) \leq F(B, \mathbf{p}, \lambda, x) \qquad \forall A, B \in \mathcal{S}^{d\times d},\ A \geq B,$$

where $A \geq B$ *means that* $A - B$ *is a nonnegative definite matrix.*

We note that when F is differentiable, the ellipticity condition can also be defined by requiring that the matrix $\frac{\partial F}{\partial D^2 u}$ is negative semi-definite (cf. [32, p. 441]).

Lastly, we will assume that fully nonlinear second order problems satisfy a comparison principle, as represented by the following definition:

Definition 1.4. *Problem* (1.3b) *is said to satisfy a* comparison principle *if the following statement holds. For any upper semi-continuous function* u *and lower semi-continuous function* v *on* $\overline{\Omega}$*, if* u *is a viscosity subsolution and* v *is a viscosity supersolution of* (1.3b) *(see Definition 1.6), then* $u \leq v$ *on* $\overline{\Omega}$*.*

Remark 1.1. *Since the comparison principle immediately infers the uniqueness of viscosity solutions, it is also called a* strong uniqueness property *for problem* (1.3b) *(cf. [3]).*

1.2.1 Viscosity Solutions

Much of the PDE theory for fully nonlinear first order problems has become well understood thanks to the viscosity solution concept pioneered by Crandall and Lions in the early 1980s. We will focus on the viscosity solution framework as it is presented in [17] and [13]. Due to the full nonlinearity in (1.3a) and (1.3b), standard weak solution theory based upon multiplication by a regular test function and integration by parts is not applicable. Furthermore, it can be shown that multiple solutions

that satisfy the PDE almost everywhere exist. Thus, a different solution concept was necessary in order to guarantee existence and uniqueness for solutions of fully nonlinear PDEs. As such, the following definition of viscosity solutions for first order problems was proposed (see [13]):

Definition 1.5. *Let H denote the first order operator in (1.3a).*

(i) A locally bounded function $u : (\Omega \cup \Gamma) \to \mathbb{R}$ is called a viscosity subsolution *of (1.3a) if $\forall \varphi \in C^1(\Omega \cup \Gamma)$, when $u^* - \varphi$ has a local maximum at $x_0 \in \Omega \cup \Gamma$,*

$$H_*(\nabla\varphi(x_0), u^*(x_0), x_0) \leq 0.$$

(ii) A locally bounded function $u : (\Omega \cup \Gamma) \to \mathbb{R}$ is called a viscosity supersolution *of (1.3a) if $\forall \varphi \in C^1(\Omega \cup \Gamma)$, when $u_* - \varphi$ has a local minimum at $x_0 \in \Omega \cup \Gamma$,*

$$H^*(\nabla\varphi(x_0), u_*(x_0), x_0) \geq 0.$$

(iii) A locally bounded function $u : (\Omega \cup \Gamma) \to \mathbb{R}$ is called a viscosity solution *of (1.3a) if u is both a viscosity subsolution and viscosity supersolution of (1.3a).*

The definition captures intrinsic properties of viscosity solutions that were historically linked to the vanishing viscosity method presented in Section 1.2.2.

The definition of viscosity solutions can be extended to fully nonlinear second order partial differential equations. Thus, we have the following definition of a viscosity solution for second order problems that readily extends Definition 1.5:

Definition 1.6. *Assume the second order operator F in (1.3b) is elliptic in a function class $\mathcal{A} \subset B(\Omega)$.*

(i) The function $u \in \mathcal{A}$ is called a viscosity subsolution *of (1.3b) if $\forall \varphi \in C^2(\overline{\Omega})$, when $u^* - \varphi$ has a local maximum at $x_0 \in \overline{\Omega}$,*

$$F_*(D^2\varphi(x_0), \nabla\varphi(x_0), u^*(x_0), x_0) \leq 0.$$

(ii) *The function $u \in \mathcal{A}$ is called a* viscosity supersolution *of* (1.3b) *if* $\forall \varphi \in C^2(\overline{\Omega})$, *when $u_* - \varphi$ has a local minimum at $x_0 \in \overline{\Omega}$,*

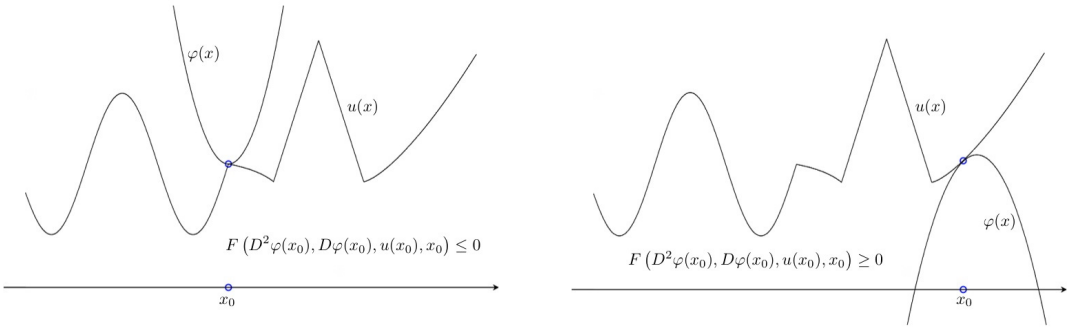
$$F^*(D^2\varphi(x_0), \nabla\varphi(x_0), u_*(x_0), x_0) \geq 0.$$

(iii) *The function $u \in \mathcal{A}$ is called a* viscosity solution *of* (1.3b) *if u is both a viscosity subsolution and a viscosity supersolution of* (1.3b).

We note that the ellipticity assumption on the operator F in the definition is not necessary. However, the ellipticity assumption is used when proving the existence of viscosity solutions.

The above definitions can be informally interpreted as follows. Without a loss of generality, we may assume $u^*(x_0) = \varphi(x_0)$ whenever $u^* - \varphi$ has a local maximum at x_0 or $u_*(x_0) = \varphi(x_0)$ whenever $u_* - \varphi$ has a local minimum at x_0. Then, u is a viscosity solution of (1.3b) if for all smooth functions φ such that φ "touches" the graph of u^* from above at x_0, we have $F_*(D^2\varphi(x_0), \nabla\varphi(x_0), \varphi(x_0), x_0) \leq 0$, and for all smooth functions φ such that φ "touches" the graph of u_* from below at x_0, we have $F^*(D^2\varphi(x_0), \nabla\varphi(x_0), \varphi(x_0), x_0) \geq 0$. For the first order problem represented by (1.3a), the interpretation remains the same with F replaced by H and the Hessian term $D^2\varphi(x_0)$ removed from the evaluation of the operator. Assuming $u^* = u_*$, i.e. u is continuous, the informal geometric interpretation of a viscosity solution for second order problems in one dimension is pictured in Figure 1.1.

We now make a few observations based on Definitions 1.5 and 1.6. The first observation is that if u and H (or F) are continuous, then the upper and lower $*$ indices can be removed. The second observation is that both definitions are nonvariational. The solution concept is not based upon an integration by parts approach. Instead, the viscosity solution concept is based upon a local "differentiation by parts" approach that characterizes an intrinsic property of viscosity solutions. The third observation is that for a general fully nonlinear PDE problem, uniqueness of viscosity solutions may only hold in a restricted function class. While the viscosity solution concept can



(a) A viscosity subsolution. **(b)** A viscosity supersolution.

Figure 1.1: A geometric interpretation of viscosity solutions for second order problems.

eliminate "generalized" solutions that satisfy the PDE problem almost everywhere, a viscosity solution may not exist without the added ellipticity assumption. In general, fully nonlinear second order operators are not globally elliptic, and ellipticity only holds in a restricted function class. Furthermore, fully nonlinear first order PDEs are degenerate in the sense that the operators H and $-H$ are both elliptic using Definition 1.3 due to the fact H is independent of D^2u.

We end this section by mentioning that the viscosity solution concept is readily extended to dynamic PDE problems of the form

$$u_t + H(\nabla u, u, x, t) = 0 \qquad \text{in } (\Omega \cup \Gamma) \times (0, \infty), \tag{1.4a}$$

$$v_t + F(D^2 v, \nabla v, v, x, t) = 0 \qquad \text{in } \overline{\Omega} \times (0, \infty), \tag{1.4b}$$

complemented with an initial condition. Then, for Hamilton-Jacobi problems, we have $u : (\Omega \cup \Gamma) \times (0, \infty) \to \mathbb{R}$ is a viscosity solution on $(\Omega \cup \Gamma) \times (0, T]$ if it satisfies the given initial condition and for every $\varphi \in C^1\big((\Omega \cup \Gamma) \times (0, \infty)\big)$ and $T > 0$:

1. If (x_0, t_0) is a local maximum point of $u^* - \varphi$ on $(\Omega \cup \Gamma) \times (0, T]$, then

$$u_t(x_0, t_0) + H_*(\nabla\varphi(x_0, t_0), u^*(x_0, t_0), x_0, t_0) \leq 0.$$

2. If (x_0, t_0) is a local minimum point of $u_* - \varphi$ on $(\Omega \cup \Gamma) \times (0, T]$, then

$$u_t(x_0, t_0) + H^*(\nabla\varphi(x_0, t_0), u_*(x_0, t_0), x_0, t_0) \geq 0.$$

The extension for a fully nonlinear second order operator F is analogous.

1.2.2 Existence and Uniqueness

We now provide some background material for the analysis of fully nonlinear PDEs. First, we present a classical existence and uniqueness result. Then, we introduce the vanishing viscosity method, which provides a constructive way to build viscosity solutions for first order fully nonlinear PDEs. Finally, we reference some of the major contributions that led to a full existence and uniqueness theory for second order problems.

We first consider the Cauchy problem associated with the Hamilton-Jacobi equation, (1.4a), paired with the initial condition $u(x, 0) = u_0(x)$ for all $x \in \mathbb{R}^d$. Suppose the PDE operator H is given by $H(\nabla u)$. Then, it was first shown by Crandall and Lions in [14] that when $H \in C(\mathbb{R}^d)$ and $u_0 \in BUC(\mathbb{R}^d)$, there exists a unique function $u \in BUC(\mathbb{R}^d \times [0, T])$, for all $T > 0$, such that u is the unique viscosity solution of (1.4a) and $u(x, 0) = u_0(x)$. The existence and uniqueness result can also be extended to the Dirichlet boundary condition form of (1.4a) and/or when H depends on u, x, and t.

Historically, viscosity solutions for Hamilton-Jacobi equations (1.4a) were obtained using a limiting procedure. Consider the following family of second order quasi-linear Cauchy problems

$$u_t^\epsilon + H(\nabla u^\epsilon, u^\epsilon, x) - \epsilon\Delta u^\epsilon = 0, \tag{1.5}$$

where $\epsilon \in (0, 1)$. Then, it was also first proven by Crandall and Lions in [17] that there exists a unique solution u^ϵ to (1.5) for each ϵ, and the family of solutions converges locally and uniformly to a continuous function u that was shown to be exactly the

viscosity solution of the Cauchy version of (1.4a) as characterized by Definition 1.5. Constructing a sequence of solutions to a regularized version of (1.3a) as a means to approximate the viscosity solution is referred to as the *vanishing viscosity method*, see Section 1.3.1.

In contrast to the quick success of the viscosity solution theory for first order problems, the theory for second order problems took longer to develop. The first evidence of a general existence and uniqueness theory can be found in Lions' works [41] and [42] for Hamilton-Jacobi-Bellman equations (see Section 1.4.2). The major breakthrough for proving uniqueness and existence of solutions was the development of a maximum-principle by Jensen in [37] and the development of Perron's method by Ishii (see [35] and [15]). In general, a limiting procedure for constructing viscosity solutions for second order problems has not been established. However, some limiting procedures have been obtained such as those of Evans found in [21], and the idea of constructing solutions to a regularized problem forms the basis of the vanishing moment method of Feng and Neilan to be discussed in Section 1.3.1. A self-contained overview of the basic theory of viscosity solutions for fully nonlinear second order PDEs can be found in [16].

1.3 Overview of Numerical Methods for Fully Nonlinear PDEs

We now present a brief overview of some of the existing methods for numerically approximating viscosity solutions for fully nonlinear second order PDEs. We start with describing some of the numerical difficulties that arise when approximating fully nonlinear PDEs. Next, we classify and list some of the successful approximation techniques. We note that while a few numerical methods have been developed for second order fully nonlinear PDEs, limited tools currently exist for the convergence

analysis of these methods. A more expansive overview of numerical methods for second order fully nonlinear PDEs can be found in the survey paper [27].

By the definition of viscosity solutions, we see that an approximation method must be able to capture low-regularity functions. However, to make the situation more difficult, an approximation method must also have a mechanism for filtering the possibly infinite number of lower-regularity functions that satisfy the PDE almost everywhere whenever the viscosity solution has higher regularity. We will see that such low-regularity almost everywhere solutions can correspond to algebraic solutions of the system of equations that results from discretizing a fully nonlinear PDE. These false algebraic solutions are referred to as *numerical artifacts* resulting from the discretization, and these numerical artifacts are known to plague the numerical discretization of fully nonlinear PDEs when using or adapting standard numerical methods for linear, semi-linear, and quasi-linear PDEs. Additionally, viscosity solutions may be unique only in a restrictive function class $\mathcal{A}$. Consequently, the numerical solutions must also belong to a discrete function class that is consistent with $\mathcal{A}$.

A good example of numerical artifacts can be found in [27], where the Monge-Ampère equation (see Section 1.4.1) in two-dimensions with a C^∞ solution is approximated on the unit square with a standard nine-point FD scheme. By using a Newton solver and varying the initial guess, the authors demonstrate the ability to capture all $2^{(N-2)^2}$ numerical solutions when using an $N \times N$ grid with $N = 4$. Yet, only one of the solutions corresponds to the C^∞ convex solution of the PDE.

Due to the nonlinearity of the PDE, multiplication by a test function and using integration by parts is not possible. In fact, the definition of viscosity solutions is entirely nonvariational. Thus, Galerkin based methodologies are not immediately applicable for approximating fully nonlinear PDEs. Instead, the viscosity solution concept is based on a differentiation by parts approach, an entirely local definition that has no known discrete analogue. For first order problems, the vanishing viscosity method provides a constructive proof of existence for viscosity solutions

that can be utilized to design numerical approximation schemes. However, a complete constructive theory does not currently exist for second order problems.

1.3.1 Indirect Methods

A natural starting point for approximating viscosity solutions is to mimic the original concept of viscosity solutions based upon limiting procedures for "stabilized" equations. Therefore, the first class of methods that we introduce are all indirect methods. The methods are based on approximating the given PDE problem by another PDE that is in turn discretized. Thus, the approximation error has two components, the PDE approximation and the numerical approximation of the latter problem. The first two methods listed are based on approximating fully nonlinear PDEs with higher-order quasi-linear PDEs. The third method is based on transforming a PDE problem into a constrained optimization problem that can be approximated in a least-squares sense.

We first state the classical result of Crandall and Lions found in [14] for viscosity solutions of Hamilton-Jacobi equations. Consider the Cauchy problem using the Hamilton-Jacobi equation given by (1.4a) and the second order quasi-linear PDE (1.5). Assume the operator H is locally Lipschitz in $\mathbb{R}^d$, $H = H(\nabla u)$, and the initial data u_0 is bounded and Lipschitz continuous in $\mathbb{R}^d$ with $T > 0$. Then, if u^ϵ denotes the solution to (1.5) and u denotes the viscosity solution of (1.4a), there holds

$$\sup_{0\le t\le T}\sup_{x\in\mathbb{R}^d}\left|u^\epsilon(x,t)-u(x,t)\right| \le c\sqrt{\epsilon},$$

where c depends only on T and the Lipschitz constants of u_0 and H. Furthermore the Lipschitz continuity assumption for H is only used to guarantee u^ϵ is smooth. If H is only continuous and $u_t^\epsilon, \Delta u^\epsilon \in L^p_{\text{loc}}\left(\mathbb{R}^d \times (0,\infty)\right)$ for $1 \le p < \infty$, then the estimate still holds.

The result of Crandall and Lions provides the foundation of the *vanishing viscosity method* for approximating first order fully-nonlinear PDEs. In general, the vanishing

viscosity method for approximating the viscosity solution of (1.3a) is defined as approximating the viscosity solution u by a numerical approximation for the second order quasi-linear PDE

$$-\epsilon \Delta u^{\epsilon} + H\left(\nabla u^{\epsilon}, u^{\epsilon}, x\right) = 0 \qquad \text{in } \Omega \cup \Gamma,$$

where ϵ is a small positive constant. Numerically, the perturbed second order problem should have better properties at the discrete level due to the less severe nature of the nonlinearity. Historically, the nomenclature for vanishing viscosity comes from relating the second-order term Δu^{ϵ} to the viscosity tensor from continuum mechanics.

Building upon the success of the vanishing viscosity method for first order problems, Feng and Neilan first proposed the *vanishing moment method* for fully nonlinear second order PDEs in [29]. In full generality, the vanishing moment method is defined as approximating (1.1) with the perturbed equation

$$G_{\epsilon}\left[u^{\epsilon}\right] + F\left[u^{\epsilon}\right] = 0 \qquad \text{in } \Omega, \qquad \epsilon > 0, \tag{1.6}$$

where G_{ϵ} is a differential operator with order greater than two. Furthermore, the following criteria were proposed for the choice of G_{ϵ}:

(i) G_{ϵ} should be a linear or quasi-linear operator.

(ii) $-G_{\epsilon}$ should be an elliptic operator.

(iii) G_{ϵ} should "vanish" in some sense as $\epsilon \to 0$.

(iv) Equation (1.6) should be "easy" to solve numerically.

In essence, the higher order operator G_{ϵ} should be chosen such that the "weak" solution u^{ϵ} converges to the viscosity solution of (1.1) in an appropriate norm as $\epsilon \to 0$.

In practice, the vanishing moment method has typically been applied with

$$G_\epsilon[v] := \epsilon \Delta^2 v,$$

where Δ^2 denotes the fourth order biharmonic operator. When $d = 2$ and u^ϵ is interpreted as the vertical displacement of a bent plate, then the moment tensor of the plate $D^2 u^\epsilon$ can be associated with weak forms of $\Delta^2 u^\epsilon$, motivating the terminology "vanishing moment". An additional boundary condition such as

$$\Delta u^\epsilon = \epsilon \qquad \text{or} \qquad \frac{\partial \Delta u^\epsilon}{\partial n} = \epsilon \qquad \text{or} \qquad D^2 u^\epsilon n \cdot n = \epsilon \qquad \text{on } \partial\Omega$$

is typically chosen to complement the Dirichlet boundary condition so that the perturbed problem is well-posed.

The vanishing moment method has typically been used in concert with Galerkin type methods for discretizing (1.6), and it has been applied to problems such as the Monge-Ampère equation, the equation of prescribed Gauss curvature, and the infinity-Laplace equation, which is a quasi-linear PDE with non-divergence form. Much of the analysis for the vanishing moment method is in relation to the Monge-Ampère equation. A general convergence result such as that of Crandall and Lions for first order problems is still open for the vanishing moment method.

The third example of indirect methods comes from the least-squares approach of Dean and Glowinski as found in [19] and [20]. The main idea is to transform the fully nonlinear second order boundary value problem into a constrained optimization problem of the form

$$j(u, p) \leq j(v, q) \qquad \forall \{v, q\} \in s,$$

for an appropriate functional j and an appropriate set s. One example is to use the least squares functional for j; hence, we are led to the least squares method. The approximation corresponds to the solution of a set of normal equations related to

the optimization problem. Such a methodology has been considered for the Monge-Ampère equation and the Pucci equation. A general overview of such methods can be found in [27]. We note that least-squares methods will not be further considered in this dissertation.

1.3.2 Direct Methods

In this section, we provide references for a non-exhaustive sampling of various direct numerical methods for approximating solutions of fully nonlinear first and second order PDEs that are most relevant to the numerical methods presented in the dissertation. We first consider methods for Hamilton-Jacobi equations. Then, we will discuss some FD methods for second order equations. We end with references for alternative Galerkin-type methods for second order equations.

We begin by mentioning two numerical methods for Hamilton-Jacobi equations that can be considered analogous to the methods we develop for second order fully nonlinear equations. The first is the FD methods of Crandall and Lions as defined in [14]. The paper defines a set of sufficient conditions that, when satisfied, guarantee convergence of a FD method to the underlying viscosity solution. One of the key ideas is the use of a numerical Hamiltonian. Later, Yan and Osher extended the use of numerical Hamiltonians to nonstandard local DG methods as a means to develop higher-order schemes in [53]. More details about the Crandall and Lions framework can be found in Chapter 2, and more details about the Yan and Osher DG methods can be found in Chapter 3.

A seminal work concerning the design of convergent FD methods for Hamilton-Jacobi problems is the result of Tadmor concerning hyperbolic conservation laws. In summary, every convergent monotone FD scheme for Hamilton-Jacobi equations as well as hyperbolic conservation laws must contain a numerical diffusion/viscosity term (cf. [51]). Thus, direct converging monotone methods for nonlinear first order problems implicitly approximate the original differential equation with a perturbation

term that involves an approximation for a second order operator such as the Laplacian operator. Consequently, the vanishing viscosity method and the approximation of viscosity solutions are strongly correlated.

Following the extension of viscosity solution theory to fully nonlinear second order problems, the foundational paper concerning the numerical analysis of second order problems was [3] by Barles and Souganidis. Their paper provides a set of sufficient conditions that guarantee convergence for a class of approximation methods. However, one of the first known methods to satisfy the framework was the wide-stencil FD method of Oberman developed nearly twenty years later for approximating the Monge-Ampère equation (see [46]). We do note that many FD methods which do not necessarily fulfill the Barles and Souganidis framework have been developed for Hamilton-Jacobi-Bellman equations (cf. [4], [38], [27], and the references therein). In general, direct approximations using Galerkin methods are limited. The existing methods assume the viscosity solution is actually a classical solution, such as in [5]. Methods for Hamilton-Jacobi-Bellman equations that do not have regularity requirements typically enforce a structure requirement on the second order components of the linear operators, as in [36]. Thus, while some progress has been made, numerical methods for fully nonlinear second order problems are not nearly as rich as the numerical methods for Hamilton-Jacobi problems.

1.4 Selected Applications of Second Order Fully Nonlinear PDEs

Second order fully nonlinear PDEs have applications in many science and engineering fields such as antenna design, astrophysics, differential geometry, fluid mechanics, image processing, meteorology, mesh generation, optimal control, optimal mass transport, etc (see [27] and the references therein). As computing technology increases and the field of computational science continues to grow as a new means for

scientific research and practical problem-solving, the necessity for numerical methods to approximate and visualize solutions to PDE problems has become paramount. In this section, we present two of the prototypical second order fully nonlinear PDEs that arise in many applications. The first is the Monge-Ampère equation and the second is the Hamilton-Jacobi-Bellman equation. These two prototypical fully nonlinear PDEs will serve as the basis for testing the numerical methods developed in the dissertation.

1.4.1 Monge-Ampère Equations

We first introduce the Monge-Ampère equation from differential geometry. Let $u : \Omega \to \mathbb{R}$ be a continuous function, and define the subdifferential of u at x_0 by

$$\partial u(x_0) := \{p \mid u(x) \geq u(x_0) + p \cdot (x - x_0) \forall x \in \Omega\}.$$

Let $\partial u(E) := \bigcup_{x \in E} \partial u(x)$ for all $E \subset \Omega$. Then, the Monge-Ampère measure associated with u is defined by

$$\mathcal{M}_u(E) := \mathcal{L}^d\big(\partial u(E)\big) \qquad \forall \text{ Borel sets } E \subset \Omega,$$

where $\mathcal{L}^d$ denotes the Lebesgue measure on $\mathbb{R}^d$. The Monge-Ampère problem then involves finding a continuous convex function u with given Dirichlet boundary data such that $\mathcal{M}_u = \mu$ for a given Radon measure μ.

Suppose

$$\int_E f d\mathcal{L}^d = \mu(E)$$

for some function $f : \Omega \to \mathbb{R}$. It can be shown that if μ is absolutely continuous with respect to $\mathcal{L}^d$ and $u \in C^2(\overline{\Omega})$, then there holds

$$\det D^2 u = f \tag{1.7}$$

pointwise. To see this, use the fact $\partial u(x_0) = \nabla u(x_0)$ for $u \in C^1(\overline{\Omega})$ and Sard's Theorem ([44]) to obtain

$$\int_E f d\mathcal{L}^d = \mu(E) = \mathcal{M}_u(E) = \int_{\nabla u(E)} d\mathcal{L}^d = \int_E \det\,(D^2 u)\, d\mathcal{L}^d$$

for all Borel sets $E \subset \Omega$. Equation (1.7) is referred to as the Monge-Ampère equation. A more complete derivation can be found in [47] and [48].

A solution u that satisfies the weakened form of the Monge-Ampère equation $\mathcal{M}_u = \mu$ and given Dirichlet data is called an Aleksandrov solution to the Monge-Ampère equation. In general, for a non-strictly convex domain Ω, the Monge-Ampère equation may not have a classical solution even if the source f, Dirichlet boundary data, and $\partial\Omega$ are smooth (cf. [32]). However, Aleksandrov solutions exist uniquely when $f > 0$ (cf. [2]). Furthermore, the Aleksandrov solution is also a viscosity solution provided μ is absolutely continuous with respect to $\mathcal{L}^d$ and has a continuous density f. In fact, when $f > 0$, the two solution concepts are equivalent (cf. [33]). We do note that other continuous nonconvex solutions of (1.7) with given Dirichlet data may exist even when $f > 0$, and the Monge-Ampère operator is only elliptic in the class of convex functions (cf. [32]).

Besides its applications in differential geometry, the Monge-Ampère operator also arises in other application areas such as in Riemannian geometry and in optimal mass transport. For instance, the equation of prescribed Gauss curvature from Riemannian geometry involves the Monge-Ampère operator. Suppose a hypersurface of $\mathbb{R}^{d+1}$ is the graph of some function u such that, at each point of the surface, the Gauss curvature equals a prescribed constant K. Then, we have u satisfies the second order fully nonlinear PDE

$$\det D^2 u = K\left(1 + |\nabla u|^2\right)^{(d+2)/2},$$

which is called the prescribed Gauss curvature equation. The Monge-Kantorovich optimal transport equation is another PDE that involves the Monge-Ampère operator.

Let two sets $X_1, X_2 \subset \mathbb{R}^d$, with mass density functions f_1, f_2, respectively, have equal mass. Then, the optimal mass-preserving mapping between the two sets, u, subject to a given positive quadratic cost density involves the fully nonlinear second order PDE constraint equation

$$\det D^2 u = \frac{f_1}{f_2}.$$

More information can be found in [52].

1.4.2 Hamilton-Jacobi-Bellman Equations

We now introduce Hamilton-Jacobi-Bellman (HJB) equations. We will focus on how they relate to stochastic optimal control with regards to the Bellman principle, an approach for transforming a stochastic optimal control problem into a second order fully nonlinear PDE problem. Thus, approximating solutions of HJB equations provides a means for approximating solutions to stochastic optimal control problems. We will first explain how a special form of HJB equations is related to stochastic optimal control. We end the section with the statement of the general form of HJB equations. A more detailed introduction to stochastic optimal control and HJB equations can be found in [30] and [31].

We begin with a generic stochastic optimal control problem. Let $t, T \in \mathbb{R}$ and $T > t$. Furthermore, let $\mathbf{X} : [t, T] \to \mathbb{R}^d$; $f : [t, T] \times \mathbb{R}^d \times A \to \mathbb{R}^d$, for $A \subset \mathbb{R}^n$; and $\sigma : [t, T] \times \mathbb{R}^d \times A \to \mathbb{R}^d$. Then, we assume the stochastic process $\mathbf{X}(\tau)$ is governed by the stochastic differential equation

$$d\mathbf{X}(\tau) = f\big(\tau, \mathbf{X}(\tau), \mathbf{a}(\tau)\big) + \sigma\big(\tau, \mathbf{X}(\tau), \mathbf{a}(\tau)\big)\, dW(\tau) \qquad \forall \tau \in (t, T] \tag{1.8a}$$

$$\mathbf{X}(t) = x, \tag{1.8b}$$

where W is a Wiener process, $\mathbf{a} : (t, T] \to A$ is a control vector, and each function also depends upon a stochastic variable w that has been suppressed. Thus, we have the control vector $\mathbf{a}$ determines the state of the stochastic process $\mathbf{X}$ at each

time τ through the stochastic differential equation, where the initial conditions are determined by t and x.

We introduce a cost functional $J : (0, T] \times \mathbb{R}^n \times A \to \mathbb{R}$ defined by

$$J(t, x, \mathbf{a}) := E_{t,x}\left[\int_t^T L\big(\tau, \mathbf{X}(\tau), \mathbf{a}(\tau)\big)d\tau + g\big(\mathbf{X}(T)\big)\right], \tag{1.9}$$

where L is called a running cost, g is called a terminal cost, and $E_{t,x}$ denotes the expected value functional. Then, the stochastic optimal control problem involves finding the optimal control $\mathbf{a}^*(t)$ such that the cost functional J is minimized for all t, i.e.,

$$J(t, \mathbf{X}^*, \mathbf{a}^*) = \min_{\mathbf{a}\in A} J(t, \mathbf{X}, \mathbf{a}), \tag{1.10}$$

where $\mathbf{X}(t)$ is the solution to (1.8) corresponding to $\mathbf{a}(t)$ at each time t.

We now apply dynamic programming to convert the stochastic optimal control problem (1.10) into a second order fully nonlinear PDE problem. Suppose $\mathbf{a}^* \in A$ such that

$$\mathbf{a}^* \in \operatorname*{argmin}_{\mathbf{a}\in A} J(t, x, \mathbf{a}),$$

and define the value function $V : \mathbb{R}^d \times (0, T] \to \mathbb{R}$ by

$$V(x, t) := J(t, x, \mathbf{a}^*). \tag{1.11}$$

Then, V is the minimal cost achieved starting from the initial state $\mathbf{X}(t) = x$ at time t, and $\mathbf{a}^*$ is the optimal control that attains the minimum.

Let $\Omega \subset \mathbb{R}^d$ and $T > 0$. Then, the Bellman principle states that the value function V satisfies the second order fully nonlinear PDE

$$\frac{\partial V}{\partial t} = \inf_{\mathbf{a}\in A}\big(L_{\mathbf{a}}[V] - h_{\mathbf{a}}\big) \qquad \text{in } \Omega \times (0, T], \tag{1.12}$$

where

$$L_{\mathbf{a}}[V] := \frac{1}{2}\sigma_{\mathbf{a}} \otimes \sigma_{\mathbf{a}} : D^2V + b_{\mathbf{a}} \cdot \nabla V,$$

with

$$\sigma_{\mathbf{a}} := \sigma\big(t, x, \mathbf{a}(t)\big), \qquad b_{\mathbf{a}} := f\big(t, x, \mathbf{a}(t)\big), \qquad h_{\mathbf{a}} := L\big(t, x, \mathbf{a}(t)\big).$$

Notationally, $\otimes$ denotes the outer product for two vectors and $B : C$ denotes the Frobenius inner product for two matrices $B, C \in \mathbb{R}^{d\times d}$. We note that the PDE defined by (1.12) is an instance of the more general HJB equation which will be defined at the end of the section.

Supposing V is known, we have

$$\mathbf{a}^* = \underset{\mathbf{a}\in A}{\operatorname{argmin}}\big(L_{\mathbf{a}}V - h_{\mathbf{a}}\big), \tag{1.13a}$$

$$d\mathbf{X}^*(\tau) = f\big(\tau, \mathbf{X}^*(\tau), \mathbf{a}^*(\tau)\big) + \sigma\big(\tau, \mathbf{X}^*(\tau), \mathbf{a}^*(\tau)\big)\, dW(\tau) \qquad \forall \tau \in (t, T] \tag{1.13b}$$

with $\mathbf{X}^*(t) = x$. Thus, given V, we can find the optimal control and corresponding stochastic process that solve the stochastic optimal control problem. Therefore, solving the HJB equation provides a means for solving the stochastic optimal control problem (1.10).

The preceding formulation of the HJB equation applies when the HJB equation comes from a stochastic optimal control problem. In general, the HJB problem does not necessarily correspond to a stochastic optimal control problem. Let $\Theta \subset \mathbb{R}^m$ and $\Omega_T := \Omega \times (0, T]$. Suppose $A^\theta : \Omega_T \to \mathbb{R}^{d\times d}$, $b^\theta : \Omega_T \to \mathbb{R}^d$, and $c^\theta, f_\theta : \Omega_T \to \mathbb{R}$. Then, we define the HJB equation by

$$u_t = \inf_{\theta\in\Theta}\big(L_\theta u - f_\theta\big), \tag{1.14}$$

where

$$L_\theta u := A^\theta : D^2u + b^\theta \cdot \nabla u + c^\theta\, u$$

and

$$A^\theta : D^2 u := \sum_{i=1}^{d} \sum_{j=1}^{d} a_{i,j}^\theta \, u_{x_i x_j}, \qquad b^\theta \cdot \nabla u := \sum_{i=1}^{d} b_i^\theta \, u_{x_i}.$$

Thus, the HJB equation involves taking an infimum over a family of second order linear differential operators. Since the optimal value for θ may change at each point in the domain Ω_T, the solution u will correspond to a function $\theta^* : \Omega_T \to \Theta$ that specifies a particular second order linear operator at each point in the domain. Throughout the dissertation, we refer to (1.14) and do not specify a corresponding stochastic optimal control problem if it exists whenever we refer to the HJB equation.

1.5 Dissertation Organization

The dissertation is organized as follows. In Chapter 2 we introduce a new FD framework for second order fully nonlinear elliptic partial differential equations. We show that the proposed FD methods converge to the underlying viscosity solution for one-dimensional problems, and we provide examples of such methods. We also extend the framework to parabolic problems and elliptic problems in higher dimensions. In Chapter 3 we present a DG framework based upon a nonstandard local DG formulation that is shown to naturally generalize the methods proposed in Chapter 2. Thus, the methods in Chapter 3 provide a way to increase the accuracy of the FD methods first introduced in Chapter 2. To complement and expand upon the methods of Chapter 3, we present an alternative DG methodology based upon a nonstandard interior-penalty DG formulation in Chapter 4. All numerical methods presented in Chapters 2, 3, and 4 are direct methods. In Chapter 5 we propose an indirect methodology for approximating Hamilton-Jacobi-Bellman equations that incorporates the vanishing moment method. Chapters 2, 3, 4, and 5 all include numerical experimental results that support the proposed methodologies. Finally, in Chapter 6, we comment on open problems and future directions for extending the work of the dissertation.

We note that many of the topics discussed in the dissertation can be found in previously submitted works. The material in Chapter 2 builds upon the one-dimensional FD framework first presented in [28] for elliptic problems. In particular, numerical tests are added as well as extensions for lower-order terms, higher-dimensional problems, and parabolic problems. The material in Chapter 3 expands upon the one-dimensional work of [23] using results that can be found in [26] and [40]. Additional numerical tests are provided as well as an extended treatment of high-dimensional problems. Chapter 4 is directly based upon [25] and [24]. However, the material of Chapter 5 is appearing for the first time in this dissertation.

1.6 Mathematical Software and Implementation

We end the chapter with a comment regarding the numerical test data found throughout the dissertation. All of the numerical results in Chapters 2, 3, and 4 were produced using code developed in the programming language Matlab ([43]). The use of specific Matlab functions is documented in the relevant sections of the dissertation. All of the numerical results in Chapter 5 were produced using the finite element method software package COMSOL. More specifically, the numerical test data in Section 5.3.1 was produced using version 3.5a ([11]), and the numerical test data in Section 5.3.2 was produced using version 4.0a ([12]) in conjunction with Matlab through the LiveLink feature. More information on Matlab can be found at http://www.mathworks.com, and more information on COMSOL can be found at http://www.comsol.com. The experiments in Section 5.3.1 were run on a laptop with an Intel Core Duo processor rated at 2.0 GHz. All of the other experiments were run on a laptop with an Intel Core i5 processor rated at 2.53 GHz.

Chapter 2

Finite Difference Methods

In this chapter we develop a general and practical framework for building convergent finite difference (FD) methods for approximating viscosity solutions of second order Dirichlet boundary value problems

$$F[u](x) := F\left(D^2u(x), \nabla u(x), u(x), x\right) = 0, \qquad x \in \Omega \subset \mathbb{R}^d, \tag{2.1a}$$

$$u(x) = g(x), \qquad x \in \partial\Omega, \tag{2.1b}$$

where F is a fully nonlinear elliptic operator and Ω is an open, bounded domain. After developing a set of sufficient conditions under which a given class of FD methods converges, we will provide examples of methods that satisfy the proposed conditions. Then we extend the framework to second order parabolic problems. Numerical tests that demonstrate the convergence of the proposed methods for elliptic and parabolic problems will be presented. We will also provide the corresponding numerical PDE theory for fully nonlinear first order PDE problems for motivation and comparison throughout the chapter.

2.1 Difference Operators

To construct FD methods for problem (2.1), we need to introduce difference operators for approximating first and second order derivatives. To this end, we first form a computational grid for the domain Ω. For simplicity, we will assume Ω is a d-rectangle, i.e., $\Omega = (a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_d, b_d)$, and we shall only consider grids that are uniform in each coordinate x_i, $i = 1, 2, \ldots, d$. Let J_i be a positive integer and $h_i = \frac{b_i - a_i}{J_i - 1}$ for $i = 1, 2, \ldots, d$. Define $h = (h_1, h_2, \ldots, h_d) \in \mathbb{Z}^d$, $J = \prod_{i=1}^d J_i$, and $\mathbb{N}_J^d = \{\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d) \mid 1 \le \alpha_i \le J_i, i = 1, 2, \ldots, d\}$. Then, $\left|\mathbb{N}_J^d\right| = J$. We divide Ω into $\prod_{i=1}^d (J_i - 1)$ subdomains with grid points

$$x_\alpha = \Big(a_1 + (\alpha_1 - 1)h_1, a_2 + (\alpha_2 - 1)h_2, \ldots, a_d + (\alpha_d - 1)h_d\Big)$$

for each multi-index $\alpha \in \mathbb{N}_J^d$. We call $\mathcal{T}_h = \{x_\alpha\}_{\alpha \in \mathbb{N}_J^d}$ a grid (set of nodes) for $\overline{\Omega}$.

Using the grid $\mathcal{T}_h$, we can define the standard forward and backward difference operators. Let $\{e_i\}_{i=1}^d$ denote the canonical basis for $\mathbb{R}^d$. Define the forward and backward difference operators by

$$\delta^+_{x_i,h_i} v(x) := \frac{v(x + h_i e_i) - v(x)}{h_i}, \qquad \delta^-_{x_i,h_i} v(x) := \frac{v(x) - v(x - h_i e_i)}{h_i} \tag{2.2}$$

for a function v defined on Ω and

$$\delta^+_{x_i,h_i} V_\alpha := \frac{V_{\alpha + e_i} - V_\alpha}{h_i}, \qquad \delta^-_{x_i,h_i} V_\alpha := \frac{V_\alpha - V_{\alpha - e_i}}{h_i}$$

for a grid function V defined on the grid $\mathcal{T}_h$. Note that "ghost-values" may need to be introduced in order for the above difference operators to be well-defined on the boundary of Ω. In the following, the operators $\delta^+_{x_i,h_i}$ and $\delta^-_{x_i,h_i}$ for $i = 1, 2, \ldots, d$ will serve as the building blocks in the construction of our FD methods in the sense that we shall approximate all first and second order derivatives by using combinations and compositions of these two operators.

To approximate $u_{x_i}(x_\alpha)$, we have two straight-forward options; that is,

$$u_{x_i}(x_\alpha) \approx \delta^+_{x_i,h_i} u(x_\alpha), \qquad u_{x_i}(x_\alpha) \approx \delta^-_{x_i,h_i} u(x_\alpha).$$

As a result, we have four possible ways to approximate $u_{x_i x_j}(x_\alpha)$ using strictly composition operators; that is,

$$\begin{aligned} u_{x_i x_j}(x_\alpha) &\approx \delta^+_{x_j,h_j}\delta^+_{x_i,h_i} u(x_\alpha), \quad u_{x_i x_j}(x_\alpha) \approx \delta^-_{x_j,h_j}\delta^-_{x_i,h_i} u(x_\alpha), \\ u_{x_i x_j}(x_\alpha) &\approx \delta^+_{x_j,h_j}\delta^-_{x_i,h_i} u(x_\alpha), \quad u_{x_i x_j}(x_\alpha) \approx \delta^-_{x_j,h_j}\delta^+_{x_i,h_i} u(x_\alpha). \end{aligned}$$

A main idea for approximating viscosity solutions is to take advantage of multiple approximations for a first or second order derivative in order to better capture the behavior of the target function.

We now express explicit representations for the composition operators used above to approximate second order derivatives. Let v denote a function defined on Ω and V denote a grid function defined on the grid $\mathcal{T}_h$. Choose $i, j \in \{1, 2, \ldots, d\}$. Observe,

$$\delta^+_{x_j,h_j}\delta^+_{x_i,h_i} v(x) = \frac{v(x + h_i e_i + h_j e_j) - v(x + h_i e_i) - v(x + h_j e_j) + v(x)}{h_i h_j}, \tag{2.3a}$$

$$\delta^+_{x_j,h_j}\delta^-_{x_i,h_i} v(x) = \frac{v(x + h_j e_j) - v(x) - v(x - h_i e_i + h_j e_j) + v(x - h_i e_i)}{h_i h_j}, \tag{2.3b}$$

$$\delta^-_{x_j,h_j}\delta^+_{x_i,h_i} v(x) = \frac{v(x + h_i e_i) - v(x + h_i e_i - h_j e_j) - v(x) + v(x - h_j e_j)}{h_i h_j}, \tag{2.3c}$$

$$\delta^-_{x_j,h_j}\delta^-_{x_i,h_i} v(x) = \frac{v(x - h_i e_i - h_j e_j) - v(x - h_i e_i) - v(x - h_j e_j) + v(x)}{h_i h_j} \tag{2.3d}$$

and

$$\delta^+_{x_j,h_j}\delta^+_{x_i,h_i}V_\alpha = \frac{V_{\alpha+e_i+e_j} - V_{\alpha+e_i} - V_{\alpha+e_j} + V_\alpha}{h_i h_j},$$
$$\delta^+_{x_j,h_j}\delta^-_{x_i,h_i}V_\alpha = \frac{V_{\alpha+e_j} - V_\alpha - V_{\alpha-e_i+e_j} + V_{\alpha-e_i}}{h_i h_j},$$
$$\delta^-_{x_j,h_j}\delta^+_{x_i,h_i}V_\alpha = \frac{V_{\alpha+e_i} - V_{\alpha+e_i-e_j} - V_\alpha + V_{\alpha-e_j}}{h_i h_j},$$
$$\delta^-_{x_j,h_j}\delta^-_{x_i,h_i}V_\alpha = \frac{V_{\alpha-e_i-e_j} - V_{\alpha-e_i} - V_{\alpha-e_j} + V_\alpha}{h_i h_j}.$$

The following lemma examines the local truncation errors for the given mixed second order derivative approximation operators:

Lemma 2.1. *For $i, j \in \{1, 2, \ldots, d\}$ such that $i \neq j$, the operators $\delta^{\pm}_{x_j,h_j}\delta^{\pm}_{x_i,h_i}$ and $\delta^{\mp}_{x_j,h_j}\delta^{\pm}_{x_i,h_i}$ have first order local truncation errors, and the operators $(\delta^+_{x_j,h_j}\delta^+_{x_i,h_i} + \delta^-_{x_j,h_j}\delta^-_{x_i,h_i})/2$ and $(\delta^+_{x_j,h_j}\delta^-_{x_i,h_i} + \delta^-_{x_j,h_j}\delta^+_{x_i,h_i})/2$ have second order local truncation errors.*

Proof. Suppose $v \in C^4(\Omega)$. Pick $i, j \in \{1, 2, \ldots, d\}$ such that $i \neq j$. Observe,

$$\begin{aligned}
v(x \pm h_i e_i) &= v(x) \pm h_i v_{x_i}(x) + \frac{1}{2}h_i^2 v_{x_i x_i}(x) \pm \frac{1}{6}h_i^3 v_{x_i x_i x_i}(x) + \mathcal{O}\left(h_i^4\right), \\
v(x \pm h_i e_i \pm h_j e_j) &= v(x) \pm h_i v_{x_i}(x) \pm h_j v_{x_j}(x) + \frac{1}{2}h_i^2 v_{x_i x_i}(x) + h_i h_j v_{x_i x_j}(x) \\
&\quad + \frac{1}{2}h_i^2 v_{x_i x_i}(x) \pm \frac{1}{6}h_i^3 v_{x_i x_i x_i}(x) \pm \frac{1}{2}h_i^2 h_j v_{x_i x_i x_j}(x) \\
&\quad \pm \frac{1}{2}h_i h_j^2 v_{x_i x_j x_j}(x) \pm \frac{1}{6}h_j^3 v_{x_j x_j x_j}(x) + \mathcal{O}\left(h_i^4 + h_j^4\right), \\
v(x \mp h_i e_i \pm h_j e_j) &= v(x) \mp h_i v_{x_i}(x) \pm h_j v_{x_j}(x) + \frac{1}{2}h_i^2 v_{x_i x_i}(x) - h_i h_j v_{x_i x_j}(x) \\
&\quad + \frac{1}{2}h_i^2 v_{x_i x_i}(x) \mp \frac{1}{6}h_i^3 v_{x_i x_i x_i}(x) \pm \frac{1}{2}h_i^2 h_j v_{x_i x_i x_j}(x) \\
&\quad \mp \frac{1}{2}h_i h_j^2 v_{x_i x_j x_j}(x) \pm \frac{1}{6}h_j^3 v_{x_j x_j x_j}(x) + \mathcal{O}\left(h_i^4 + h_j^4\right).
\end{aligned}$$

Thus, we have

$$\begin{aligned}\delta^{\pm}_{x_j,h_j}\delta^{\pm}_{x_i,h_i}v(x) &= \frac{v(x\pm h_ie_i\pm h_je_j)-v(x\pm h_ie_i)-v(x\pm h_je_j)+v(x)}{h_ih_j}\\ &= \frac{h_ih_jv_{x_ix_j}(x)\pm\frac{1}{2}h_i^2h_jv_{x_ix_ix_j}(x)\pm\frac{1}{2}h_ih_j^2v_{x_ix_jx_j}(x)+\mathcal{O}\left(h_i^4+h_j^4\right)}{h_ih_j}\\ &= v_{x_ix_j}(x)\pm\frac{1}{2}h_iv_{x_ix_ix_j}(x)\pm\frac{1}{2}h_jv_{x_ix_jx_j}(x)+\mathcal{O}\left(h_i^2+h_j^2\right)\end{aligned}$$

and

$$\begin{aligned}\delta^{\pm}_{x_j,h_j}\delta^{\mp}_{x_i,h_i}v(x) &= \frac{v(x\pm h_je_j)-v(x)-v(x\mp h_ie_i\pm h_je_j)+v(x\mp h_ie_i)}{h_ih_j}\\ &= \frac{h_ih_jv_{x_ix_j}(x)\mp\frac{1}{2}h_i^2h_jv_{x_ix_ix_j}(x)\pm\frac{1}{2}h_ih_j^2v_{x_ix_jx_j}(x)+\mathcal{O}\left(h_i^4+h_j^4\right)}{h_ih_j}\\ &= v_{x_ix_j}(x)\mp\frac{1}{2}h_iv_{x_ix_ix_j}(x)\pm\frac{1}{2}h_jv_{x_ix_jx_j}(x)+\mathcal{O}\left(h_i^2+h_j^2\right),\end{aligned}$$

and the result follows. The proof is complete. □

The standard central difference operators for approximating first and second order derivatives in one-dimension can also be expressed in terms of the various difference operators defined above. Choose $i \in \{1, 2, \ldots, d\}$. The standard central difference operator for approximating first order derivatives in one-dimension is defined by

$$\delta_{x_i,h_i}v(x) := \frac{v(x+h_ie_i)-v(x-h_ie_i)}{2h_i} \tag{2.4}$$

for a function v on Ω and

$$\delta_{x_i,h_i}V_\alpha := \frac{V_{\alpha+e_i}-V_{\alpha-e_i}}{2h_i}$$

for a grid function V on the grid $\mathcal{T}_h$. The standard central difference operator for approximating second order derivatives in one-dimension is defined by

$$\delta^2_{x_i,h_i} v(x) := \frac{v(x + h_i e_i) - 2v(x) + v(x - h_i e_i)}{h_i^2} \tag{2.5}$$

for a function v on Ω and

$$\delta^2_{x_i,h_i} V_\alpha := \frac{V_{\alpha+e_i} - 2V_\alpha + V_{\alpha-e_i}}{h_i^2}$$

for a grid function V on the grid $\mathcal{T}_h$. Then, a simple computation reveals

$$\delta_{x_i,h_i} v(x) = \frac{1}{2}\left(\delta^+_{x_i,h_i} + \delta^-_{x_i,h_i}\right) v(x)$$

and

$$\begin{aligned}
&\delta^+_{x_i,h_i}\delta^+_{x_i,h_i} v(x) = \delta^2_{x_i,h_i} v(x + h_i e_i), && \delta^+_{x_i,h_i}\delta^-_{x_i,h_i} v(x) = \delta^2_{x_i,h_i} v(x),\\
&\delta^-_{x_i,h_i}\delta^+_{x_i,h_i} v(x) = \delta^2_{x_i,h_i} v(x), && \delta^-_{x_i,h_i}\delta^-_{x_i,h_i} v(x) = \delta^2_{x_i,h_i} v(x - h_i e_i)
\end{aligned}$$

for a function v defined on Ω and

$$\delta_{x_i,h_i} V_\alpha = \frac{1}{2}\left(\delta^+_{x_i,h_i} + \delta^-_{x_i,h_i}\right) V_\alpha$$

and

$$\begin{aligned}
&\delta^+_{x_i,h_i}\delta^+_{x_i,h_i} V_\alpha = \delta^2_{x_i,h_i} V_{\alpha+e_i}, && \delta^+_{x_i,h_i}\delta^-_{x_i,h_i} V_\alpha = \delta^2_{x_i,h_i} V_\alpha,\\
&\delta^-_{x_i,h_i}\delta^+_{x_i,h_i} V_\alpha = \delta^2_{x_i,h_i} V_\alpha, && \delta^-_{x_i,h_i}\delta^-_{x_i,h_i} V_\alpha = \delta^2_{x_i,h_i} V_{\alpha-e_i}
\end{aligned}$$

for a grid function V defined on $\mathcal{T}_h$. Thus, we have $\delta^+_{x_i,h_i}\delta^-_{x_i,h_i}$ and $\delta^-_{x_i,h_i}\delta^+_{x_i,h_i}$ yield the standard central difference operator in one dimension, $\delta^+_{x_i,h_i}\delta^+_{x_i,h_i}$ yields the standard central difference operator in one dimension shifted forwards one grid point in the x_i

coordinate direction, and $\delta^-_{x_i,h_i}\delta^-_{x_i,h_i}$ yields the standard central difference operator in one dimension shifted backwards one grid point in the x_i coordinate direction.

Lemma 2.2. *For $i \in \{1, 2, \ldots, d\}$, the operator $\delta^{\pm}_{x_i,h_i}\delta^{\pm}_{x_i,h_i}$ has first order local truncation error, and the operators $\delta^{\pm}_{x_i,h_i}\delta^{\mp}_{x_i,h_i}$ and $(\delta^+_{x_i,h_i}\delta^+_{x_i,h_i} + \delta^-_{x_i,h_i}\delta^-_{x_i,h_i})/2$ have second order local truncation errors.*

Proof. Suppose $v \in C^4(\Omega)$. Pick $i \in \{1, 2, \ldots, d\}$. Observe,

$$v(x \pm 2h_i e_i) = v(x) \pm 2h_i v_{x_i}(x) + 2h_i^2 v_{x_i x_i}(x) \pm \frac{4}{3}h_i^3 v_{x_i x_i x_i}(x) + \mathcal{O}(h_i^4).$$

Thus, we have

$$\begin{aligned}
\delta^{\pm}_{x_i,h_i}\delta^{\pm}_{x_i,h_i}v(x) &= \delta^2_{x_i,h_i}v(x \pm h_i e_i) \\
&= \frac{v(x \pm 2h_i e_i) - 2v(x \pm h_i e_i) + v(x)}{h_i^2} \\
&= \frac{h_i^2 v_{x_i x_i}(x) \pm \frac{2}{3}h_i^3 v_{x_i x_i x_i}(x) + \mathcal{O}(h_i^4)}{h_i^2} \\
&= v_{x_i x_i}(x) \pm \frac{2}{3}h_i v_{x_i x_i x_i}(x) + \mathcal{O}(h_i^2)
\end{aligned}$$

and

$$\begin{aligned}
\delta^{\pm}_{x_i,h_i}\delta^{\mp}_{x_i,h_i}v(x) &= \delta^2_{x_i,h_i}v(x) \\
&= \frac{v(x + h_i e_i) - 2v(x) + v(x - h_i e_i)}{h_i^2} \\
&= \frac{h_i^2 v_{x_i x_i}(x) + \mathcal{O}(h_i^4)}{h_i^2} \\
&= v_{x_i x_i}(x) + \mathcal{O}(h_i^2),
\end{aligned}$$

and the result follows. The proof is complete. □

The standard central difference operators for approximating second order mixed derivatives can also be formed using the second order composition operators defined

by (2.3). Choose $i, j \in \{1, 2, \ldots, d\}$ such that $i \neq j$. The standard central difference operator for approximating second order mixed derivatives is defined by

$$\begin{aligned}\delta^2_{x_i,x_j;h_i,h_j} v(x) := &\frac{v(x + h_i e_i + h_j e_j) - v(x - h_i e_i + h_j e_j)}{4h_i h_j} \\ &+ \frac{v(x - h_i e_i - h_j e_j) - v(x + h_i e_i - h_j e_j)}{4h_i h_j}\end{aligned} \tag{2.6}$$

for a function v on Ω and

$$\delta^2_{x_i,x_j;h_i,h_j} V_\alpha := \frac{V_{\alpha+e_i+e_j} - V_{\alpha-e_i+e_j} - V_{\alpha+e_i-e_j} + V_{\alpha-e_i-e_j}}{4h_i h_j}$$

for a grid function V on the grid $\mathcal{T}_h$. Then, a simple computation reveals

$$\begin{aligned}\delta^2_{x_i,x_j;h_i,h_j} &= \delta_{x_i,h_i}\delta_{x_j,h_j} \\ &= \frac{\delta^+_{x_i,h_i}\delta^+_{x_j,h_j} + \delta^+_{x_i,h_i}\delta^-_{x_j,h_j} + \delta^-_{x_i,h_i}\delta^+_{x_j,h_j} + \delta^-_{x_i,h_i}\delta^-_{x_j,h_j}}{4}.\end{aligned}$$

Thus, the standard central difference operator for approximating second order mixed derivatives is formed by averaging two second order operators, and we immediately have the following lemma using the results of Lemma 2.1:

Lemma 2.3. *For $i, j \in \{1, 2, \ldots, d\}$ such that $i \neq j$, the operator $\delta^2_{x_i,x_j;h_i,h_j}$ has second order local truncation error.*

Using the above difference operators, we can define discrete gradient and Hessian operators. To this end, using (2.2) and (2.3), we define forward and backward discrete gradients by

$$\nabla^\pm_h := \left[\delta^\pm_{x_1,h_1}, \delta^\pm_{x_2,h_2}, \cdots, \delta^\pm_{x_d,h_d}\right]^T, \tag{2.7}$$

and four discrete Hessians using composition of the two discrete gradients by

$$D^{+\pm}_h := \nabla^+_h \left(\nabla^\pm_h\right)^T, \qquad D^{-\pm}_h := \nabla^-_h \left(\nabla^\pm_h\right)^T, \tag{2.8}$$

where T denotes the transpose operation. Both of the discrete gradients defined by (2.7) are essential building blocks in the Crandall and Lions FD framework for first order Hamilton-Jacobi problems, as discussed in the following section. Likewise, the discrete Hessians defined by (2.8) will all be used as building blocks for our FD framework for second order fully nonlinear PDE problems. We also note that the standard second order discrete gradient and Hessian operators formed by using (2.4) and (2.5) are defined by

$$\nabla_h := [\delta_{x_1,h_1}, \delta_{x_2,h_2}, \cdots, \delta_{x_d,h_d}]^T \tag{2.9}$$

and

$$\left[D_h^2\right]_{i,j} := \begin{cases} \delta^2_{x_i,h_i}, & \text{if } i = j, \\ \delta^2_{x_i,h_i;x_j,h_j}, & \text{otherwise,} \end{cases} \tag{2.10}$$

respectively. Furthermore, using Lemmas 2.1, 2.2, and 2.3, we have the following alternative second order discrete Hessian operators

$$\overline{D}_h^2 := \frac{D_h^{+-} + D_h^{-+}}{2}, \qquad \widetilde{D}_h^2 := \frac{D_h^{++} + D_h^{--}}{2}, \tag{2.11}$$

and

$$\left[\widehat{D}_h^2\right]_{i,j} := \begin{cases} \frac{\delta^+_{x_i,h_i}\delta^+_{x_i,h_i} + \delta^-_{x_i,h_i}\delta^-_{x_i,h_i}}{2}, & \text{if } i = j, \\ \delta^2_{x_i,h_i;x_j,h_j}, & \text{otherwise} \end{cases} \tag{2.12}$$

which will also be considered in the following.

2.2 The Finite Difference Framework of Crandall and Lions for Hamilton-Jacobi Problems

As a prelude to presenting our framework for fully nonlinear second order problems, we first present the successful FD framework of Crandall and Lions for fully nonlinear

first order Hamilton-Jacobi problems as found in [14]. For this section, we consider approximating the viscosity solution $u \in \mathcal{A} \subset C^0\big(\Omega \times (0,T]\big)$ for the Hamilton-Jacobi problem

$$u_t + H\left(\nabla u\right) = 0 \qquad \text{in } \Omega \times (0,T], \tag{2.13a}$$

$$u = g \qquad \text{on } \Gamma \subset \partial\Omega \times (0,T], \tag{2.13b}$$

$$u = u_0 \qquad \text{on } \Omega \times \{0\}, \tag{2.13c}$$

where the operator H is a continuous and possibly nonlinear function, $\mathcal{A}$ is a function class in which the unique viscosity solution u resides, and $\Omega \subset \mathbb{R}^d$. We note that the following framework can also be adapted for H a function of ∇u, u, x, and t.

The FD framework will be described for two-dimensional problems, i.e., $d = 2$. Pick an integer $N > 0$ and let $\Delta t = T/N$. We use a super-index to denote the approximation at a given time level. Thus, $U^n_{j,k}$ denotes the approximation for $u(x_{j,k}, n\Delta t)$, $n = 0, 1, \ldots, N$. The schemes to be considered will all be explicit in time and can be written in the form

$$U^{n+1}_{j,k} = G\left(U^n_{j-p,k-r}, \ldots, U^n_{j+q+1,k+s+1}\right), \tag{2.14}$$

where p, q, r, s are fixed nonnegative integers and G is a function of $(p+q+2)(r+s+2)$ variables.

Definition 2.1.

(i) *The FD scheme represented by* (2.14) *is said to have* differenced form *if there exists a function g such that*

$$\begin{aligned} &G\left(U_{j-p,k-r}, \ldots, U_{j+q+1,k+s+1}\right) \\ &\quad = U_{j,k} - \Delta t\, g\Big(\delta^+_{x_1,h_1} U_{j-p,k-r}, \ldots, \delta^+_{x_1,h_1} U_{j+q,k+s+1}; \\ &\qquad\qquad \delta^+_{x_2,h_2} U_{j-p,k-r}, \ldots, \delta^+_{x_2,h_2} U_{j+q+1,k+s}\Big). \end{aligned} \tag{2.15}$$

(ii) *The FD scheme represented by* (2.14) *is said to be* consistent *with* (2.13a) *if*

$$g(a,\dots,a;b,\dots,b) = H\left([a,b]^T\right) \tag{2.16}$$

for all $a, b \in \mathbb{R}$.

(iii) *The FD scheme represented by* (2.14) *is said to be* monotone *on* $[-R,R]$ *if the function* $G\left(U_{j-p,k-r},\dots,U_{j+q+1,k+s+1}\right)$ *is nondecreasing in each argument whenever* $\left|\delta^+_{x_1,h_1}U_{\ell,m}\right|, \left|\delta^+_{x_2,h_2}U_{\ell',m'}\right| \le R$ *for* $j-p \le \ell \le j+q$, $k-r \le m \le k+s+1$, $j-p \le \ell' \le j+q+1$, $k-r \le m' \le k+s$.

(iv) *The function* g *is called a* numerical Hamiltonian *for the FD scheme represented by* (2.14) *when* (2.15) *holds.*

Using Definition 2.1, Crandall and Lions proved the following convergence result (adapted to the case $d = 2$):

Theorem 2.1. *Let* $H : \mathbb{R}^2 \to \mathbb{R}$ *be continuous and* u_0 *be bounded and Lipschitz continuous on* $\mathbb{R}^2$ *with* L *as a Lipschitz constant. For* $\Delta t/h_1, \Delta t/h_2 > 0$ *fixed, let the FD scheme* (2.14) *have differenced form, be monotone on* $[-(L+1), L+1]$, *and be consistent with* (2.13a). *Assume the numerical Hamiltonian* g *is locally Lipschitz continuous. Define* U^0 *by* $U^0_{j,k} := u_0(x_{j,k})$ *and* U^n, $n = 1, 2, \dots, N$, *by* (2.14). *Let* u *be the viscosity solution of* (2.13). *Then there is a constant* c *depending only on* $\sup|u_0|$, L, g, *and* T *such that*

$$\left|U^n_{j,k} - u\left(x_{j,k}, n\Delta t\right)\right| \le c\sqrt{\Delta t}$$

for $0 \le n \le N$ *and all* j, k.

Remark 2.1.

(a) Using the backward difference operators and fixing values for p,q,r, and s, we have schemes with the form

$$G(U_{j,k}) = U_{j,k} - \Delta t\, g\left(\delta^+_{x_1,h_1} U_{j,k}, \delta^-_{x_1,h_1} U_{j,k}; \delta^+_{x_2,h_2} U_{j,k}, \delta^-_{x_2,h_2} U_{j,k}\right)$$

are of differenced form. Using vector notation, we have

$$G(U_{j,k}) = U_{j,k} - \Delta t\, g\left(\nabla^+_h U_{j,k}, \nabla^-_h U_{j,k}\right).$$

Then, the monotonicity requirement implies g is nonincreasing with respect to $\nabla^+_h U_{j,k}$ *and nondecreasing with respect to* $\nabla^-_h U_{j,k}$*; that is,* $g(\downarrow, \uparrow)$.

(b) The idea of using multiple derivative approximations and requiring monotonicity in each approximation will also be vitally used in our FD framework for fully nonlinear second order problems to be developed in Sections 2.3, 2.4, and 2.5.

We now present two families of numerical Hamiltonians that satisfy the structure conditions of the Crandall and Lions FD framework (cf. [49] and the references therein). The first example is the Lax-Friedrichs numerical Hamiltonian $\widehat{H}_{LF}$ defined by

$$\widehat{H}_{LF}\left(q^+, q^-, v, x\right) := H\left(\frac{q^+ + q^-}{2}, v, x\right) - \beta \cdot \left(q^+ - q^-\right) \tag{2.17}$$

for β an undetermined positive constant or function that enforces the monotonicity of $\widehat{H}$. The second example is the Godunov numerical Hamiltonian $\widehat{H}_G$ defined by

$$\widehat{H}_G\left(q^+, q^-, v, x\right) := \operatorname{ext}_{q_1 \in I(q_1^+, q_1^-)} \operatorname{ext}_{q_2 \in I(q_2^+, q_2^-)} \cdots \operatorname{ext}_{q_d \in I(q_d^+, q_d^-)} H(q, v, x) \tag{2.18}$$

for

$$I(\alpha, \beta) = [\min\{\alpha, \beta\}, \max\{\alpha, \beta\}]$$

and the function ext defined by

$$\text{ext}_{v\in I(\alpha,\beta)} = \begin{cases} \max_{\alpha\le v\le\beta} & \text{if } \alpha \le \beta, \\ \min_{\beta\le v\le\alpha} & \text{if } \alpha > \beta. \end{cases} \tag{2.19}$$

Remark 2.2.

(a) The term $-\beta\cdot(q^+ - q^-)$ *in* $\widehat{H}_{LF}$ *is called a* numerical viscosity *due to the fact*

$$-\sum_{i=1}^{d}\left(\delta^+_{x_i,h_i}U_j - \delta^-_{x_i,h_i}U_j\right) = -h\,\frac{U_{j-1}-2U_j+U_{j+1}}{h^2} = -h\,\delta^2_{x_i,h_i}U_j,$$

a central difference approximation of $u_{xx}(x_j)$ *scaled by* h*. Thus, Lax-Friedrichs FD methods for Hamilton-Jacobi problems are direct realizations of the vanishing viscosity method (see Section 1.3.1).*

(b) Both the Lax-Friedrichs numerical Hamiltonian and the Godunov numerical Hamiltonian will serve as inspiration for the design of specific FD schemes in our FD framework for fully nonlinear second order problems to be developed in Sections 2.3, 2.4, and 2.5.

2.3 A New Finite Difference Framework for One-Dimensional Second Order Elliptic Problems

We now propose a new FD framework for approximating viscosity solutions of fully nonlinear second order PDEs, as represented by problem (2.1), in the special case $d = 1$. The key concepts introduced will be that of numerical operators and numerical moments, which can be considered analogous to the concepts of numerical Hamiltonians and numerical viscosities, respectively, that were used in Crandall and Lions FD framework for Hamilton-Jacobi equations presented above in Section 2.2.

2.3.1 Definitions

Based on the idea of using both the forward and backward discrete gradient operators, (2.7), in the framework for fully nonlinear first order PDE problems, we propose building a FD framework for fully nonlinear second order PDE problems that incorporates all four discrete Hessian operators given by (2.8). Recall that in one-dimension, we have

$$D_h^{++}v(x) = \delta_{x,h}^2 v(x+h), \qquad D_h^{+-}v(x) = \delta_{x,h}^2 v(x),$$
$$D_h^{-+}v(x) = \delta_{x,h}^2 v(x), \qquad D_h^{--}v(x) = \delta_{x,h}^2 v(x-h),$$

for a function v on Ω and

$$D_h^{++}V_i = \delta_{x,h}^2 V_{i+1}, \qquad D_h^{+-}V_i = \delta_{x,h}^2 V_i,$$
$$D_h^{-+}V_i = \delta_{x,h}^2 V_i, \qquad D_h^{--}V_i = \delta_{x,h}^2 V_{i+1},$$

for a grid function V on $\mathcal{T}_h = \{x_1, x_2, \ldots, x_J\}$ with $x_1 := a$ and $x_J := b$. Thus, in one-dimension, we only consider three difference approximations for second order derivatives that can all be expressed in terms of the standard central difference operator $\delta_{x,h}^2$.

The above simple argument motivates us to propose the following general FD method for equation (2.1) in one-dimension: Find a grid function U such that

$$\widehat{F}\left(\delta_{x,h}^2 U_{i-1}, \delta_{x,h}^2 U_i, \delta_{x,h}^2 U_{i+1}, \delta_{x,h}^+ U_i, \delta_{x,h}^- U_i, U_i, x_i\right) = 0 \tag{2.20}$$

for $i = 2, 3, \cdots, J-1$. As expected, U_i is intended to be an approximation of $u(x_i)$ for $i = 1, 2, \cdots, J$. Also, U_0 and U_{J+1} are two ghost values.

Definition 2.2. *The function* $\widehat{F} : \mathbb{R}^3 \times \mathbb{R}^2 \times \mathbb{R} \times \Omega \to \mathbb{R}$ *in* (2.20) *is called a* numerical operator. *FD method* (2.20) *is said to be an* admissible *scheme for problem* (2.1) *in*

one-dimension if it has at least one (grid function) solution U such that $U_1 = g(a)$ and $U_J = g(b)$.

Intuitively, $\widehat{F}$ needs to be some approximation of the differential operator F in order for scheme (2.20) to be relevant to the original PDE problem. Generally, different numerical operators $\widehat{F}$ should result in different FD methods. In Section 2.3.3 we will present two types of numerical operators that will serve as analogues to the Lax-Friedrichs numerical Hamiltonian and the Godunov numerical Hamiltonian presented in Section 2.2. For now, we propose a set of sufficient conditions that $\widehat{F}$ should satisfy in order to guarantee that the FD method proposed by (2.20) converges to the viscosity solution of problem (2.1). The conditions will be reflected in the following definition:

Definition 2.3.

(i) *Let $p, q \in \overline{\mathbb{R}}$, $v \in \mathbb{R}$, and $x \in \overline{\Omega}$. FD method (2.20) is said to be a* consistent *scheme if $\widehat{F}$ satisfies*

$$\liminf_{\substack{p_j \to p, j=1,2,3;\ q^+, q^- \to q \\ \nu \to v,\ \xi \to x}} \widehat{F}(p_1, p_2, p_3, q^+, q^-, \nu, \xi) \geq F_*(p, q, v, x), \tag{2.21a}$$

$$\limsup_{\substack{p_j \to p, j=1,2,3;\ q^+, q^- \to q \\ \nu \to v,\ \xi \to x}} \widehat{F}(p_1, p_2, p_3, q^+, q^-, \nu, \xi) \leq F^*(p, q, v, x), \tag{2.21b}$$

where F_* and F^* denote, respectively, the lower and the upper semi-continuous envelopes of F. Thus, we have

$$F_*(p, q, v, x) := \liminf_{\substack{\widetilde{p} \to p, \widetilde{q} \to q, \\ \widetilde{v} \to v, \widetilde{x} \to x}} F\big(\widetilde{p}, \widetilde{q}, \widetilde{v}, \widetilde{x}\big),$$

$$F^*(p, q, v, x) := \limsup_{\substack{\widetilde{p} \to p, \widetilde{q} \to q, \\ \widetilde{v} \to v, \widetilde{x} \to x}} F\big(\widetilde{p}, \widetilde{q}, \widetilde{v}, \widetilde{x}\big),$$

where $\widetilde{p}, \widetilde{q}, \widetilde{v} \in \mathbb{R}$ and $\widetilde{x} \in \Omega$.

(ii) *FD method* (2.20) *is said to be a* generalized monotone (g-monotone) *scheme if for each* $2 \le i \le J-1$, $\widehat{F}(p_1, p_2, p_3, q^+, q^-, v, x_i)$ *is monotone increasing in* p_1, p_3, *and* q^- *and monotone decreasing in* p_2 *and* q^+, *that is,* $\widehat{F}(\uparrow, \downarrow, \uparrow, \downarrow, \uparrow, v, x_i)$ *for* $i = 2, 3, \cdots, J-1$.

(iii) *Let* (2.20) *be an* admissible *FD method. A solution* U *of* (2.20) *is said to be* stable *if there exists a constant* $C > 0$, *which is independent of* h, *such that* U *satisfies*

$$\|U\|_{\ell^\infty(\mathcal{T}_h)} := \max_{1 \le i \le J} |U_i| \le C. \tag{2.22}$$

Also, (2.20) *is said to be a* stable scheme *if all of its solutions are stable solutions.*

Remark 2.3.

(a) The consistency and g-monotonicity (generalized monotonicity) defined above are different from those given in [3, 39, 8]. $\widehat{F}$ is asked to be monotone in $\delta_x^2 U_{j-1}, \delta_x^2 U_j$, and $\delta_x^2 U_{j+1}$, not in each individual entry U_j. To avoid confusion, we use the words "g-monotonicity" and "g-monotone" to indicate that the monotonicity is defined as above. We shall demonstrate in Section 2.3.4 that the above new definitions, especially the one for g-monotonicity, are more suitable and much easier to verify for (practical) finite difference methods. The new notions of consistency and g-monotonicity are logical extensions of their widely used counterparts for first order Hamilton-Jacobi equations, see Remark 2.1.

(b) On the other hand, the above stability definition is the same as that given in [3, 39, 8].

(c) We note that if F is a continuous function, we can also assume that $\widehat{F}$ is a continuous function. Then, (2.21a) *and* (2.21b) *reduce to the condition* $\widehat{F}(p, p, p, q, q, v, x) = F(p, q, v, x)$.

(d) *The "good" numerical operators $\widehat{F}$ we have constructed so far (cf. Section 2.3.3) all have the form*

$$\widehat{F}(p_1, p_2, p_3, q^+, q^-, \nu, \xi) = \widehat{G}(\widetilde{p}_2, p_2, q^+, q^-, \nu, \xi) \tag{2.23}$$

for some function $\widehat{G}$ and $\widetilde{p}_2 := (p_1+p_3)/2$. In other words, $\widehat{F}$ depends on p_1+p_3. Hence, a g-monotone numerical operator $\widehat{F}$ should be increasing in p_1+p_3 and decreasing in p_2, and the consistency condition reduces to

$$\liminf_{\substack{\sigma_1,\sigma_2\to p,\ q^+,q^-\to q\\ \nu\to v,\ \xi\to x}} \widehat{G}(\sigma_1, \sigma_2, q^+, q^-, \nu, \xi) \geq F_*(p, q, v, x), \tag{2.24a}$$

$$\limsup_{\substack{\sigma_1,\sigma_2\to p,\ q^+,q^-\to q\\ \nu\to v,\ \xi\to x}} \widehat{G}(\sigma_1, \sigma_2, q^+, q^-, \nu, \xi) \leq F^*(p, q, v, x). \tag{2.24b}$$

We shall need to use the above form of $\widehat{F}$ in the proof of our convergence theorem, see Theorem 2.2 below.

2.3.2 Convergence Analysis

We are now ready to state and prove a convergence theorem for FD methods defined by (2.20), which is a foundational result for this dissertation. Since the convergence will be defined locally uniformly, we first need to define a methodology for extending a given grid function to a function defined over $\overline{\Omega}$. Thus, for a given grid function U, we define a piecewise constant extension function u_h of U as follows:

$$u_h(x) := U_j \qquad \forall x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}], \quad j = 1, 2, \cdots, J, \tag{2.25}$$

where $x_{j\pm\frac{1}{2}} = x_j \pm \frac{h}{2}$ for $j = 1, 2, \cdots, J$. We now show that u_h converges to the viscosity solution of (2.1) when the underlying grid functions are defined as solutions to (2.20).

Theorem 2.2. *Suppose problem (2.1) with $d = 1$ satisfies the comparison principle of Definition 1.4 and has a unique continuous viscosity solution u. Let U be a solution to a consistent, g-monotone, and stable finite difference method (2.20) with $\widehat{F}$ satisfying (2.23), and let u_h be its piecewise constant extension given by Definition 2.25. Then u_h converges to u locally uniformly as $h \to 0^+$.*

Proof. We divide the proof into five steps.

Step 1: Since U is stable for some constant $C > 0$, it is trivial to check that u_h satisfies

$$\|u_h\|_{L^\infty(\Omega)} \leq C. \tag{2.26}$$

Define $\overline{u}, \underline{u} \in L^\infty(\Omega)$ by

$$\overline{u}(x) := \limsup_{\substack{\xi \to x \\ h \to 0^+}} u_h(\xi), \qquad \underline{u}(x) := \liminf_{\substack{\xi \to x \\ h \to 0^+}} u_h(\xi).$$

We now show that $\overline{u}$ and $\underline{u}$ are, respectively, a viscosity subsolution and a viscosity supersolution of (2.1). Hence, they must coincide by the comparison principle.

Suppose that $\overline{u} - \varphi$ takes a local maximum at $x_0 \in \Omega$ for some $\varphi \in C^2(\overline{\Omega})$. We first assume that $\varphi \in \mathbb{P}_2$, the set of all quadratic polynomials. In *Step 3* we will consider the general case $\varphi \in C^2(\overline{\Omega})$. Without loss of generality, we assume $\overline{u}(x_0) - \varphi(x_0)$ is a strict local maximum and $\overline{u}(x_0) = \varphi(x_0)$ (after a translation in the dependent variable). Then there exists a ball/interval, $B_{r_0}(x_0)$, centered at x_0 with radius $r_0 > 0$ such that

$$\overline{u}(x) - \varphi(x) < \overline{u}(x_0) - \varphi(x_0) = 0 \qquad \forall x \in B_{r_0}(x_0). \tag{2.27}$$

Thus, there exists sequences $\{h_k\}_{k \geq 1}$ and $\{\xi_k\}_{k \geq 1}$ such that, as $k \to \infty$,

$$h_k \to 0^+, \qquad \xi_k \to x_0, \qquad u_{h_k}(\xi_k) \to \overline{u}(x_0),$$
$$u_{h_k}(x) - \varphi(x) \text{ takes a local maximum at } \xi_k \text{ for sufficiently large } k,$$

and

$$\lim_{k\to\infty} \delta^2_{x,h_k} u_{h_k}(\xi_k) = \liminf_{h\to 0} \delta^2_{x,h}\overline{u}(x_0), \tag{2.28}$$

where

$$\delta^2_{x,\rho} u_h(\xi) := \frac{u_h(\xi-\rho) - 2u_h(\xi) + u_h(\xi+\rho)}{\rho^2} \quad \forall \xi \in (a+\rho, b-\rho),\ \rho > 0.$$

We remark that the right-hand side of (2.28) could either be finite or negative infinite. Then, there exists $k_0 >> 1$ such that $h_k < r_0$ and

$$0 \overset{k\to\infty}{\longleftarrow} u_{h_k}(\xi_k) - \varphi(\xi_k) \geq u_{h_k}(x) - \varphi(x) \quad \forall x \in B_{r_0}(x_0),\ k \geq k_0. \tag{2.29}$$

Step 2: Let $x = a$ denote the left endpoint of Ω and $x = b$ denote the right endpoint of Ω. Since U satisfies (2.20) with $\widehat{F}$ being of the form (2.23) at every interior grid point, it is easy to check that for $x \in \Omega_h := (a + \frac{3h}{2}, b - \frac{3h}{2})$,

$$\begin{aligned} 0 &= \widehat{F}\big(\delta^2_{x,h}u_h(x-h), \delta^2_{x,h}u_h(x), \delta^2_{x,h}u_h(x+h), \delta^+_{x,h}u_h(x), \delta^-_{x,h}u_h(x), x\big) \\ &= \widehat{G}\big(\tilde{\delta}^2_{x,h}u_h(x), \delta^2_{x,h}u_h(x), \delta^+_{x,h}u_h(x), \delta^-_{x,h}u_h(x), u_h(x), x\big), \end{aligned} \tag{2.30}$$

where

$$\tilde{\delta}^2_{x,h}u_h(x) := \big(\delta^2_{x,h}u_h(x-h) + \delta^2_{x,h}u_h(x+h)\big)/2.$$

Since $u_{h_k}(x) - \varphi(x)$ takes a local maximum at ξ_k and $h_k < r_0$ for $k \geq k_0$, by (2.29) we have

$$\delta^+_{x,h_k} u_{h_k}(\xi_k) \leq \delta^+_{x,h_k}\varphi(\xi_k), \qquad \delta^-_{x,h_k} u_{h_k}(\xi_k) \geq \delta^-_{x,h_k}\varphi(\xi_k), \tag{2.31}$$

and

$$\delta^2_{x,h_k} u_{h_k}(\xi_k) \leq \delta^2_{x,h_k}\varphi(\xi_k) = \varphi_{xx}(x_0) \tag{2.32}$$

for all $k \geq k_0$. Also, by (2.27), we get

$$\delta^2_{x,h}\overline{u}(x_0) \leq \delta^2_{x,h}\varphi(x_0) = \varphi_{xx}(x_0) \qquad \forall h \leq r_0.$$

Thus,

$$\limsup_{h\to 0} \delta^2_{x,h}\overline{u}(x_0) \le \varphi_{xx}(x_0). \tag{2.33}$$

Next, a direct computation yields that

$$\tilde{\delta}^2_{x,h}u_h(x) = \delta^2_{x,h}u_h(x) + 2R_h u_h(x), \tag{2.34}$$

where

$$R_h u_h(x) := \delta^2_{x,2h}u_h(x) - \delta^2_{x,h}u_h(x).$$

By (2.28) and the definition of lim inf we get

$$\begin{aligned}\liminf_{k\to\infty} \delta^2_{x,2h_k}u_{h_k}(\xi_k) &= \liminf_{k\to\infty} \delta^2_{x,2h_k}\overline{u}(x_0) \\ &\ge \liminf_{h\to 0} \delta^2_{x,h}\overline{u}(x_0) = \lim_{k\to\infty} \delta^2_{x,h_k}u_{h_k}(\xi_k).\end{aligned} \tag{2.35}$$

Thus,

$$\liminf_{k\to\infty} R_{h_k}u_{h_k}(\xi_k) = \liminf_{k\to\infty} \delta^2_{x,2h_k}u_{h_k}(\xi_k) - \lim_{k\to\infty} \delta^2_{x,h_k}u_{h_k}(\xi_k) \ge 0, \tag{2.36}$$

and there exists a sequence $\{\epsilon_k\}_{k\ge 1}$ and a constant $k_1 >> 1$ such that

$$\tilde{\delta}^2_{x,h_k}u_{h_k}(\xi_k) \ge \delta^2_{x,h_k}u_{h_k}(\xi_k) + \epsilon_k, \qquad \forall k \ge k_1, \tag{2.37a}$$

$$\lim_{k\to\infty} \epsilon_k = 0 \tag{2.37b}$$

by (2.34) and (2.36).

Now, it follows from (2.30), (2.31), (2.37a), and the g-monotonicity of the numerical operator $\widehat{F}$ (or $\widehat{G}$) that for $k \geq \max\{k_0, k_1\}$,

$$\begin{aligned}
0 &= \widehat{F}\big(\delta^2_{x,h_k} u_{h_k}(\xi_k - h_k), \delta^2_{x,h_k} u_{h_k}(\xi_k), \delta^2_{x,h_k} u_{h_k}(\xi_k + h_k), \\
&\qquad \delta^+_{x,h_k} u_{h_k}(\xi_k), \delta^-_{x,h_k} u_{h_k}(\xi_k), u_{h_k}(\xi_k), \xi_k\big) \\
&= \widehat{G}(\tilde{\delta}^2_{x,h_k} u_{h_k}(\xi_k), \delta^2_{x,h_k} u_{h_k}(\xi_k), \delta^+_{x,h_k} u_{h_k}(\xi_k), \delta^-_{x,h_k} u_{h_k}(\xi_k), u_{h_k}(\xi_k), \xi_k\big) \\
&\geq \widehat{G}\big(\delta^2_{x,h_k} u_{h_k}(\xi_k) + \epsilon_k, \delta^2_{x,h_k} u_{h_k}(\xi_k), \delta^+_{x,h_k} u_{h_k}(\xi_k), \delta^-_{x,h_k} u_{h_k}(\xi_k), u_{h_k}(\xi_k), \xi_k\big) \\
&\geq \widehat{G}\big(\delta^2_{x,h_k} u_{h_k}(\xi_k) + \epsilon_k, \delta^2_{x,h_k} u_{h_k}(\xi_k), \delta^+_{x,h_k} \varphi(\xi_k), \delta^-_{x,h_k} \varphi(\xi_k), u_{h_k}(\xi_k), \xi_k\big).
\end{aligned}$$

Thus, by (2.28), (2.37b), the consistency of $\widehat{F}$ (or $\widehat{G}$), and (2.33), we get

$$\begin{aligned}
0 &= \liminf_{k\to\infty} \widehat{F}\big(\delta^2_{x,h_k} u_{h_k}(\xi_k - h_k), \delta^2_{x,h_k} u_{h_k}(\xi_k), \delta^2_{x,h_k} u_{h_k}(\xi_k + h_k), \\
&\qquad \delta^+_{x,h_k} u_{h_k}(\xi_k), \delta^-_{x,h_k} u_{h_k}(\xi_k), u_{h_k}(\xi_k), \xi_k\big) \\
&\geq \liminf_{k\to\infty} \widehat{G}\big(\delta^2_{x,h_k} u_{h_k}(\xi_k) + \epsilon_k, \delta^2_{x,h_k} u_{h_k}(\xi_k), \delta^+_{x,h_k} \varphi(\xi_k), \delta^-_{x,h_k} \varphi(\xi_k), u_{h_k}(\xi_k), \xi_k\big) \\
&\geq F_*(\lim_{k\to\infty} \delta^2_{x,h_k} u_{h_k}(\xi_k), \varphi_x(x_0), \varphi(x_0), x_0) \\
&= F_*(\liminf_{h\to 0} \delta^2_{x,h} \overline{u}(x_0), \varphi_x(x_0), \varphi(x_0), x_0) \\
&\geq F_*(\limsup_{h\to 0} \delta^2_{x,h} \overline{u}(x_0), \varphi_x(x_0), \varphi(x_0), x_0) \\
&\geq F_*(\varphi_{xx}(x_0), \varphi_x(x_0), \varphi(x_0), x_0),
\end{aligned}$$

where we have used the fact that F_* is decreasing in its first argument to obtain the last two inequalities. This is true by the definition of F_* and Definition 1.3.

Step 3: We consider the general case $\varphi \in C^2(\overline{\Omega})$ which is alluded to in *Step 2*. Recall that $\overline{u} - \varphi$ is assumed to have a local maximum at x_0. Using Taylor's formula

we write

$$\begin{aligned}\varphi(x) &= \varphi(x_0) + \varphi_x(x_0)(x-x_0) + \frac{1}{2}\varphi_{xx}(x_0)(x-x_0)^2 + o(|x-x_0|^2)\\ &:= p(x) + o(|x-x_0|^2).\end{aligned}$$

For any $\epsilon > 0$, we define the following quadratic polynomial:

$$\begin{aligned}p^\epsilon(x) &:= p(x) + \epsilon(x-x_0)^2\\ &= \varphi(x_0) + \varphi_x(x_0)(x-x_0) + \left[\epsilon + \frac{\varphi_{xx}(x_0)}{2}\right](x-x_0)^2.\end{aligned}$$

Trivially, $p^\epsilon_x(x) = \varphi_x(x_0) + [2\epsilon + \varphi_{xx}(x_0)](x-x_0)$, $p^\epsilon_{xx}(x) = 2\epsilon + \varphi_{xx}(x_0)$, and $\varphi(x) - p^\epsilon(x) = o(|x-x_0|^2) - \epsilon(x-x_0)^2 \le 0$. Thus, $\varphi - p^\epsilon$ has a local maximum at x_0, Therefore, $\overline{u} - p^\epsilon$ has a local maximum at x_0. By the result of *Step 2* we have $F_*(p^\epsilon_{xx}(x_0), p^\epsilon_x(x_0), p^\epsilon(x_0), x_0) \le 0$, that is, $F_*(2\epsilon + \varphi_{xx}(x_0), \varphi_x(x_0), \varphi(x_0), x_0) \le 0$. Taking $\liminf_{\epsilon\to 0}$ and using the lower semicontinuity of F_* we obtain $0 \ge \liminf_{\epsilon\to 0} F_*(2\epsilon + \varphi_{xx}(x_0), \varphi_x(x_0), \varphi(x_0), x_0) \ge F_*(\varphi_{xx}(x_0), \varphi_x(x_0), \varphi(x_0), x_0)$. Thus, $\overline{u}$ is a viscosity subsolution of (2.1).

Step 4: By following almost the same lines as those of Step 2 and 3, we can show that if $\underline{u} - \varphi$ takes a local minimum at $x_0 \in \Omega$ for some $\varphi \in C^2(\overline{\Omega})$, then $F^*(\varphi_{xx}(x_0), \varphi_x(x_0), \varphi(x_0), x_0) \ge 0$. Hence, $\underline{u}$ is a viscosity supersolution of (2.1).

Step 5: By the comparison principle (see Definition 1.4), we get $\overline{u} \le \underline{u}$ on Ω. On the other hand, by their definitions, we have $\underline{u} \le \overline{u}$ on Ω. Thus, $\overline{u} = \underline{u}$, which coincides with the unique continuous viscosity solution u of (2.1). The proof is complete. □

Remark 2.4. *We note that the above convergence proof would also hold for the following FD method for approximating the viscosity solution of* (2.1) *for F uniformly elliptic: Find a grid function U such that*

$$\widehat{F}\left(\delta^2_{x,h}U_i, \delta^+_{x,h}U_i, \delta^-_{x,h}U_i, U_i, x_i\right) = 0 \tag{2.38}$$

for $i = 2, 3, \cdots, J-1$, where $\widehat{F}$ is consistent and stable such that $\widehat{F}$ is monotone increasing in $\delta^-_{x,h}U_i$ and monotone decreasing in $\delta^2_{x,h}U_i$ and $\delta^+_{x,h}U_i$. However, such a FD method is known not to always work for fully nonlinear second order PDE problems where the underlying operator is elliptic for only a restrictive function class, such as the Monge-Ampère equation. We will see in Section 2.3.5 that our new FD framework performs well when applied to PDE problems where the operator is not uniformly elliptic.

2.3.3 Examples of Numerical Operators

In this section we construct two classes of practical FD methods of the form (2.20). Using the first class of methods as examples, we then go through all of the steps for verifying the assumptions of Theorem 2.2 in Section 2.3.4. In particular, we present a fixed point argument for verifying the admissibility and stability of our first class of methods.

The first family of FD methods will consist of numerical operators with the form

$$\widehat{F}_b(p_1, p_2, p_3, q^+, q^-, v, x) := F\left(b_1\, p_1 + b_2\, p_2 + b_3\, p_3, \frac{q^+ + q^-}{2}, v, x\right) + \alpha\big(p_1 - 2p_2 + p_3\big) - \beta\big(q^+ - q^-\big), \tag{2.39}$$

where $\{b_j\}_{j=1}^3$ are nonnegative constants satisfying $b_1 + b_2 + b_3 = 1$, and α and β are undetermined positive constants or functions. Upon comparison with the Lax-Friedrichs numerical Hamiltonian (2.17), we refer to this class of numerical operators

as Lax-Friedrichs-like. Some specific examples from this family are

$$\widehat{F}_1(p_1,p_2,p_3,q^+,q^-,v,x) := F\Big(\frac{p_1+p_2+p_3}{3},\frac{q^++q^-}{2},v,x\Big) + \alpha\big(p_1-2p_2+p_3\big)-\beta\big(q^+-q^-\big), \tag{2.40a}$$

$$\widehat{F}_2(p_1,p_2,p_3,q^+,q^-,v,x) := F\Big(p_2,\frac{q^++q^-}{2},v,x\Big) + \alpha\big(p_1-2p_2+p_3\big)-\beta\big(q^+-q^-\big), \tag{2.40b}$$

$$\widehat{F}_3(p_1,p_2,p_3,q^+,q^-,v,x) := F\Big(\frac{p_1+2p_2+p_3}{4},\frac{q^++q^-}{2},v,x\Big) + \alpha\big(p_1-2p_2+p_3\big)-\beta\big(q^+-q^-\big). \tag{2.40c}$$

The consistency, g-monotonicity, stability, and admissibility of the Lax-Friedrichs-like schemes will be addressed in Section 2.3.4.

Remark 2.5.

(a) The term $\alpha\big(p_1-2p_2+p_3\big)$ *is called a* numerical moment *due to the fact*

$$\delta^2_{x,h}U_{j-1}-2\delta^2_{x,h}U_j+\delta^2_{x,h}U_{j+1}=h^2\,\frac{U_{j-2}-4U_{j-1}+6U_j-4U_{j+1}+U_{j+2}}{h^4},$$

a central difference approximation of $u_{xxxx}(x_j)$ *scaled by* h^2*. The role of the numerical moment will be explored numerically in Section 2.3.5.*

(b) For α *a fixed positive constant and* $\beta=0$*, Lax-Friedrichs-like numerical operators yield a direct realization of the vanishing moment methodology with* $\rho=\alpha h^2$ *(see Section 1.3.1).*

(c) Since

$$-\delta^2_{x,h}U_j=-\frac{U_{j-1}-2U_j+U_{j+1}}{h^2}=-\frac{1}{h}\left(\delta^+_{x,h}U_j-\delta^-_{x,h}U_j\right),$$

we have

$$-\alpha\,p_2=-\frac{\alpha}{h}\left(q^+-q^-\right)$$

in the implementation of the Lax-Friedrichs-like schemes. Thus, for h small, the Lax-Friedrichs-like schemes are implicitly g-monotone with respect to the first order terms even when choosing α positive and $\beta = 0$. We will see in the numerical tests of Sections 2.3.5, 2.4.2, and 2.5.2 that the Lax-Friedrichs-like schemes with $\alpha > 0$ and $\beta = 0$ perform well. In fact, the Lax-Friedrichs-like schemes with $\beta \neq 0$ appear to be limited to first order convergence similar to Lax-Friedrichs schemes for Hamilton-Jacobi equations, while Lax-Friedrichs-like schemes with $\beta = 0$ appear to exhibit second order convergence when approximating viscosity solutions with higher regularity, as seen in Sections 2.4.2 and 2.5.2.

The second family of FD methods consists of Godunov-like numerical operators due to their relation with the Godunov numerical Hamiltonian (2.18). Given $p_1, p_2, p_3, q_1, q_2 \in \mathbb{R}$, let $I(p_1, p_2, p_3)$ denote the smallest interval that contains p_1, p_2 and p_3, that is,

$$I(p_1, p_2, p_3) := \big[\min\{p_1, p_2, p_3\}, \max\{p_1, p_2, p_3\}\big],$$

and let $I(q_1, q_2)$ be defined by

$$I(q_1, q_2) := [\min\{q_1, q_2\}, \max\{q_1, q_2\}].$$

Then, the first numerical operator in this family, $\widehat{F}_4$, is defined by

$$\widehat{F}_4(p_1, p_2, p_3, q^+, q^-, v, x) := \operatorname{ext}_{p \in I(p_1,p_2,p_3)} \operatorname{ext}_{q \in I(q^+,q^-)} F(p, q, v, x), \tag{2.41}$$

where

$$\operatorname{ext}_{p\in I(p_1,p_2,p_3)} := \begin{cases} \min_{p\in I(p_1,p_2,p_3)} & \text{if } p_2 > \max\{p_1,p_3\}, \\ \max_{p\in I(p_1,p_2,p_3)} & \text{if } p_2 < \min\{p_1,p_3\}, \\ \min_{p_1\le p\le p_2} & \text{if } p_1 \le p_2 \le p_3, \\ \min_{p_3\le p\le p_2} & \text{if } p_3 \le p_2 \le p_1 \end{cases} \tag{2.42}$$

and

$$\operatorname{ext}_{q\in I(q^+,q^-)} = \begin{cases} \max_{q^+\le q\le q^-} & \text{if } q^+ \le q^-, \\ \min_{q^-\le q\le q^+} & \text{if } q^+ > q^-. \end{cases}$$

The second method in this family is a slight modification of the previous scheme, and its numerical operator, $\widehat{F}_5$, is defined by

$$\widehat{F}_5(p_1,p_2,p_3,q^+,q^-,v,x) := \operatorname{extr}_{p\in I(p_1,p_2,p_3)} \operatorname{ext}_{q\in I(q^+,q^-)} F(p,q,v,x), \tag{2.43}$$

where

$$\operatorname{extr}_{p\in I(p_1,p_2,p_3)} := \begin{cases} \min_{p\in I(p_1,p_2,p_3)} & \text{if } p_2 > \max\{p_1,p_3\}, \\ \max_{p\in I(p_1,p_2,p_3)} & \text{if } p_2 < \min\{p_1,p_3\}, \\ \max_{p_2\le p\le p_3} & \text{if } p_1 \le p_2 \le p_3, \\ \max_{p_2\le p\le p_1} & \text{if } p_3 \le p_2 \le p_1. \end{cases} \tag{2.44}$$

Lemma 2.4. *Suppose the PDE operator F is continuous. Then the Godunov-like numerical operators $\widehat{F}_4$ and $\widehat{F}_5$ are consistent and g-monotone.*

Proof. We only consider $\widehat{F}_4$. The proof for $\widehat{F}_5$ is analogous. We first show $\widehat{F}_4$ is consistent. Suppose $p_1 = p_2 = p_3 = p$ and $q^+ = q^- = q$. Then, we have $I(q^+,q^-) = \{q\}$ and $I(p_1,p_2,p_3) = \{p\}$. Thus,

$$\operatorname{ext}_{\widetilde{p}\in I(p_1,p_2,p_3)} \operatorname{ext}_{\widetilde{q}\in I(q^+,q^-)} F(\widetilde{p},\widetilde{q},v,x) = F(p,q,v,x),$$

and it follows that

$$\widehat{F}_4(p_1, p_2, p_3, q^+, q^-, v, x) = F(p, q, v, x).$$

We now show $\widehat{F}_4$ is g-monotone. We only show monotonicity in p_1, p_2, and p_3. The proof of monotonicity in q^+, q^- is analogous. We have four cases:

Case 1: $p_2 > \max\{p_1, p_3\}$. Observe, slightly increasing p_1 or p_3 yields an interval $\widetilde{I} \subset I := I(p_1, p_2, p_3)$. Thus, $\min_{\widetilde{I}} F \geq \min_I F$, and we have $\widehat{F}_4$ is increasing in p_1 and p_3. Increasing p_2 yields an interval $\widehat{I} \supset I := I(p_1, p_2, p_3)$. Thus, $\min_{\widehat{I}} F \leq \min_I F$, and we have $\widehat{F}_4$ is decreasing in p_2.

Case 2: $p_2 < \min\{p_1, p_3\}$. Observe, increasing p_1 or p_3 yields an interval $\widetilde{I} \supset I := I(p_1, p_2, p_3)$. Thus, $\max_{\widetilde{I}} F \geq \min_I F$, and we have $\widehat{F}_4$ is increasing in p_1 and p_3. Slightly increasing p_2 yields an interval $\widehat{I} \subset I := I(p_1, p_2, p_3)$. Thus, $\max_{\widehat{I}} F \leq \min_I F$, and we have $\widehat{F}_4$ is decreasing in p_2.

Case 3: $p_1 \leq p_2 \leq p_3$. Clearly $\widehat{F}_4$ is constant in p_3. Thus, $\widehat{F}_4$ is increasing in p_3. Increasing p_2 yields an interval $\widehat{I} \supset I := I(p_1, p_2, p_3)$. Thus, $\min_{\widehat{I}} F \leq \min_I F$, and we have $\widehat{F}_4$ is decreasing in p_2. Suppose $p_1 < p_2$. Then, slightly increasing p_1 yields an interval $\widetilde{I} \subset I := I(p_1, p_2, p_3)$. Thus, $\min_{\widetilde{I}} F \geq \min_I F$. Suppose $p_1 = p_2$. Then, $\widehat{F}_4\big|_{p_1,p_2,p_3} = \widehat{F}_4\big|_{p_2}$, and we have $\widehat{F}_4$ is independent of p_1. Hence, $\widehat{F}_4$ is increasing in p_1.

Case 4: $p_3 \leq p_2 \leq p_1$. Clearly $\widehat{F}_4$ is constant in p_1. Thus, $\widehat{F}_4$ is increasing in p_1. Increasing p_2 yields an interval $\widehat{I} \supset I := I(p_1, p_2, p_3)$. Thus, $\min_{\widehat{I}} F \leq \min_I F$, and we have $\widehat{F}_4$ is decreasing in p_2. Suppose $p_3 < p_2$. Then, slightly increasing p_3 yields an interval $\widetilde{I} \subset I := I(p_1, p_2, p_3)$. Thus, $\min_{\widetilde{I}} F \geq \min_I F$. Suppose $p_3 = p_2$. Then, $\widehat{F}_4\big|_{p_1,p_2,p_3} = \widehat{F}_4\big|_{p_2}$, and we have $\widehat{F}_4$ is independent of p_3. Hence, $\widehat{F}_4$ is increasing in p_3. Therefore, $\widehat{F}_4$ is g-monotone. The proof is complete. □

In Section 2.3.5, the Lax-Friedrichs-like and Godunov-like numerical operators will be tested on problems with the form $F(u_{xx}, u, x) = 0$. For such problems, the given numerical operators are trivially g-monotone with respect to the u_x parameter.

Thus, we can let $\beta = 0$ for the Lax-Friedrichs-like numerical operators and explicitly fulfill the g-monotonicity requirement. For the Godunov-like numerical operators, we use the simplified definitions

$$\widehat{F}_4(p_1, p_2, p_3, v, x) := \mathrm{ext}_{p \in I(p_1,p_2,p_3)} F(p, v, x),$$
$$\widehat{F}_5(p_1, p_2, p_3, v, x) := \mathrm{extr}_{p \in I(p_1,p_2,p_3)} F(p, v, x).$$

2.3.4 Verification of the Consistency, G-Monotonicity, Stability, and Admissibility of Lax-Friedrichs-like Finite Difference Methods

In this section we use the FD methods with the Lax-Friedrichs-like numerical operator $\widehat{F}_b$, (2.39), as examples for demonstrating all of the steps needed to verify the assumptions of the convergence theorem, Theorem 2.2. We will see that the consistency and g-monotonicity conditions are easy to verify, but the verification of the admissibility and stability conditions are more involved. For simplicity, we only consider the case in which F is purely a function of u_{xx} and x, i.e, $F = F(u_{xx}, x)$, F is differentiable, and there exists a positive constant $\gamma > 0$ such that

$$0 > -1/\gamma \geq \frac{\partial F}{\partial p} \geq -\gamma. \tag{2.45}$$

Then, we have

$$\widehat{F}_b(p_1, p_2, p_3, x) := F\big(b_1 p_1 + b_2 p_2 + b_3 p_3, x\big) + \alpha\big(p_1 - 2p_2 + p_3\big),$$

where b_1, b_2, and b_3 are nonnegative constants such that $b_1 + b_2 + b_3 = 1$.

Trivially, $\widehat{F}_b(p, p, p, x) = F(p, x)$. Hence, $\widehat{F}_b$ is a consistent numerical operator for each set of b_1, b_2, and b_3 (see Remark 2.3 (c)). To verify the g-monotonicity, we

compute

$$\frac{\partial \widehat{F}_b}{\partial p_1} = b_1 \frac{\partial F}{\partial p} + \alpha, \qquad \frac{\partial \widehat{F}_b}{\partial p_2} = b_2 \frac{\partial F}{\partial p} - 2\alpha, \qquad \frac{\partial \widehat{F}_b}{\partial p_3} = b_3 \frac{\partial F}{\partial p} + \alpha.$$

Then, $\widehat{F}_b$ is g-monotone if

$$\frac{\partial \widehat{F}_b}{\partial p_1} > 0, \qquad \frac{\partial \widehat{F}_b}{\partial p_2} < 0, \qquad \frac{\partial \widehat{F}_b}{\partial p_3} > 0.$$

On noting that $\frac{\partial F}{\partial p} \leq 0$, solving the above system of inequalities yields

$$\alpha > -\max\{b_1, b_3\} \frac{\partial F}{\partial p}. \tag{2.46}$$

Thus, we have proved the following theorem.

Theorem 2.3. *$\widehat{F}_b$ is g-monotone provided that*

$$\alpha > \max\{b_1, b_3\}\, \gamma \tag{2.47}$$

for γ defined by (2.45).

Next, we verify the admissibility and stability of the Lax-Friedrichs-like schemes. To this end, we consider the mapping $\mathcal{M}_\rho : U \to \widetilde{U}$ defined by

$$\delta_x^2 \widetilde{U}_j = \delta_x^2 U_j + \rho \widehat{F}_b\big(\delta_x^2 U_{j-1}, \delta_x^2 U_j, \delta_x^2 U_{j+1}, x_j\big), \quad j = 2, 3, \cdots, J-1. \tag{2.48}$$

Let $\mathbf{U} := (U_2, U_3, \cdots, U_{J-1})^T$ and $\widetilde{\mathbf{U}} := (\widetilde{U}_2, \widetilde{U}_3, \cdots, \widetilde{U}_{J-1})^T$. Then (2.48) can be rewritten in vector form as

$$A\widetilde{\mathbf{U}} = A\mathbf{U} + \rho \mathbf{G}(\mathbf{U}), \tag{2.49}$$

where A stands for the tridiagonal matrix corresponding to the difference operator $\delta_x^2 U_j$ and $\mathbf{G}(\mathbf{U}) = \big(G_2(\mathbf{U}, x_2), G_3(\mathbf{U}, x_3), \cdots, G_{J-1}(\mathbf{U}, x_{J-1})\big)^T$ with

$$G_j(\mathbf{U}, x_j) = \widehat{F}_b\big(\delta_x^2 U_{j-1}, \delta_x^2 U_j, \delta_x^2 U_{j+1}, x_j\big), \quad j = 2, 3, \cdots, J-1.$$

$\mathcal{M}_\rho$ is said to be *monotone* if $\widetilde{\mathbf{U}}$ is increasing in each component of $\mathbf{U}$.

Proposition 2.1. *Suppose that $\widehat{F}_b$ is g-monotone, that is, (2.47) holds. Then the mapping $\mathcal{M}_\rho$ is monotone for sufficiently small $\rho > 0$.*

Proof. Consider the following system

$$W_j = \delta_x^2 U_j, \quad j = 2, 3, \cdots, J-1, \tag{2.50}$$

$$\widetilde{W}_j = W_j + \rho \widehat{F}_b(W_{j-1}, W_j, W_{j+1}, x_j), \quad j = 2, 3, \cdots, J-1, \tag{2.51}$$

$$\delta_x^2 \widetilde{U}_j = \widetilde{W}_j, \quad j = 2, 3, \cdots, J-1. \tag{2.52}$$

Let $\mathcal{M}^{(1)} : U \to W$, $\mathcal{M}_\rho^{(2)} : W \to \widetilde{W}$, and $\mathcal{M}^{(3)} : \widetilde{W} \to \widetilde{U}$. Then, it is easy to verify that $\mathcal{M}_\rho$ can be written as a composition operator of $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$, and $\mathcal{M}^{(3)}$; that is, $\mathcal{M}_\rho := \mathcal{M}^{(3)} \circ \mathcal{M}_\rho^{(2)} \circ \mathcal{M}^{(1)}$.

Since A is positive definite, so is A^{-1}. Thus, both $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(3)}$ are monotone in the sense that they preserve the natural ordering of $\ell^\infty(\mathcal{T}_h)$. Moreover, since

$$\frac{\partial \widetilde{W}_j}{\partial W_{j-1}} = \rho \frac{\partial \widehat{F}_b}{\partial p_1}, \qquad \frac{\partial \widetilde{W}_j}{\partial W_j} = 1 + \rho \frac{\partial \widehat{F}_b}{\partial p_2}, \qquad \frac{\partial \widetilde{W}_j}{\partial W_{j+1}} = \rho \frac{\partial \widehat{F}_b}{\partial p_3},$$

then the g-monotonicity of $\widehat{F}_b$ implies that

$$\frac{\partial \widetilde{W}_j}{\partial W_{j-1}} > 0, \qquad \frac{\partial \widetilde{W}_j}{\partial W_{j+1}} > 0, \qquad \text{and} \qquad \frac{\partial \widetilde{W}_j}{\partial W_j} > 0$$

provided that

$$0 < \rho < [2\alpha + b_2/\gamma]^{-1}. \tag{2.53}$$

Thus, $\mathcal{M}_\rho^{(2)}$ is monotone, and we have $\mathcal{M}_\rho := \mathcal{M}^{(3)} \circ \mathcal{M}_\rho^{(2)} \circ \mathcal{M}^{(1)}$ is monotone, provided ρ satisfies (2.53). The proof is complete. □

Theorem 2.4. *Under the assumptions of Proposition 2.1, the FD scheme (2.20) with $\widehat{F} = \widehat{F}_b$ is admissible and stable.*

Proof. By the definition of $\mathbf{G}(\mathbf{U})$, we immediately have $\mathbf{G}(\mathbf{U} + \lambda) = \mathbf{G}(\mathbf{U})$ for any constant λ. Hence, $\mathcal{M}_\rho(\mathbf{U} + \lambda) = \mathcal{M}_\rho(\mathbf{U}) + \lambda$, and we have $\mathcal{M}_\rho$ commutes with the addition of constants. Together with the monotonicity of $\mathcal{M}_\rho$, it follows that $\mathcal{M}_\rho$ is nonexpansive in $\ell^\infty(\mathcal{T}_h)$ (cf. [18]). Hence, (2.22) holds with $C = \max\{|u_a|, |u_b|\}$, and we have the scheme is stable.

To prove admissibility of the scheme, let

$$\delta_x^2 \widetilde{V}_j = \delta_x^2 V_j + \rho \widehat{F}_b\big(\delta_x^2 V_{j-1}, \delta_x^2 V_j, \delta_x^2 V_{j+1}, x_j\big), \quad j = 2, 3, \cdots, J-1. \tag{2.54}$$

Subtracting (2.54) from (2.48) and using the mean value theorem we get

$$\begin{aligned}\delta_x^2(\widetilde{U}_j - \widetilde{V}_j) = \Big[1 + \rho \frac{\partial \widehat{F}_b}{\partial p_2}\Big]\, \delta_x^2(U_j - V_j) + \rho\, \frac{\partial \widehat{F}_b}{\partial p_1}\, \delta_x^2(U_{j-1} - V_{j-1}) \\ + \rho\, \frac{\partial \widehat{F}_b}{\partial p_3}\, \delta_x^2(U_{j+1} - V_{j+1}).\end{aligned} \tag{2.55}$$

Hence,

$$\|\widetilde{\mathbf{U}} - \widetilde{\mathbf{V}}\|_{\ell^\infty} \le \big(1 + \rho\,[(b_1 + b_3)\gamma - 1/\gamma]\,\big)\|\mathbf{U} - \mathbf{V}\|_{\ell^\infty} \le \frac{1}{2}\|\mathbf{U} - \mathbf{V}\|_{\ell^\infty}, \tag{2.56}$$

when $(b_1 + b_3) < 1/\gamma^2$ and $\rho \ge \frac{1}{2}\big[1/\gamma - (b_1 + b_3)\gamma\big]^{-1}$. Thus, (2.56) implies that the mapping $\mathcal{M}_\rho$ is contractive. By the fixed point theorem we conclude that $\mathcal{M}_\rho$ has a unique fixed point U, which in turn is the unique solution to the FD scheme (2.20) with $\widehat{F} = \widehat{F}_b$. The proof is complete. □

Remark 2.6. *We note that the choice $b_1 = b_3 = 0$ and $b_2 = 1$ trivially satisfies all of the restrictions in the proofs for any $\alpha > 0$.*

2.3.5 Numerical Experiments

We provide a series of numerical tests to demonstrate the accuracy and the order of convergence for the various proposed numerical schemes in this section. As before, we assume a uniform grid. We use the Matlab built-in nonlinear solver *fsolve* in all tests, and, unless otherwise stated, we fix the initial guess $U^{(0)}$ as the linear interpolant of the boundary data. Also, all errors are measured in the ℓ^∞ norm defined on the grid as a means to quantify the locally uniform convergence.

All tests use the numerical operators proposed in Section 2.3.3. For most tests, we record the results using $\widehat{F}_1$ and $\widehat{F}_4$. Unless otherwise stated, the results for all of the proposed Lax-Friedrichs-like numerical operators are analogous and the results for all of the proposed Godunov-like numerical operators are analogous, even though the analysis that prompts Remark 2.6 suggests $\widehat{F}_2$ could be considered preferable to $\widehat{F}_1$ and $\widehat{F}_3$. Most of the examples yield quadratic rates of convergence to the viscosity solution for the Lax-Friedrichs-like schemes. For both classes of numerical operators we observe the lack of numerical artifacts that are known to plague the standard FD discretization for fully nonlinear problems (see Section 1.3). However, for the Godunov-like schemes, this phenomena typically presents itself through the fact that the nonlinear solver *fsolve* fails to find a root. Thus, while both classes of schemes support the selectivity of the discretizations, the resulting nonlinear algebraic system appears to be better suited for *fsolve* when using the Lax-Friedrichs-like operators.

We begin with a simple power nonlinearity that has a C^∞ solution.

Example 2.1. *Consider the problem*

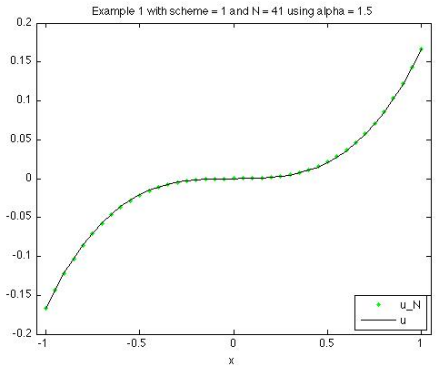
$$\begin{aligned} -u_{xx}^3 + x^3 &= 0, \qquad -1 < x < 1, \\ u(-1) = -1/6, \quad u(1) &= 1/6, \end{aligned}$$

with the exact solution $u(x) = \frac{x^3}{6}$.

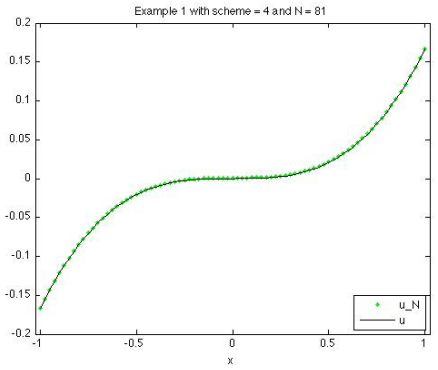
Using the linear interpolant of the boundary data as our initial guess and approximating u with the various schemes above, we obtain the computed results of Table 2.1 and Figure 2.1.

Table 2.1: Rates of convergence for Example 2.1 using $\widehat{F}_1$ and $\widehat{F}_4$ with the standard initial guess.

	$\widehat{F}_1$, $\alpha = 1.5$		$\widehat{F}_4$	
h	error	order	error	order
1.0000e-01	2.71e-02		6.40e-02	
5.0000e-02	5.10e-03	2.41	6.40e-02	0.00
2.5000e-02	1.03e-03	2.31	6.40e-02	0.00
1.2500e-02	2.33e-04	2.14	1.07e-03	5.90
6.2500e-03	5.58e-05	2.06	2.12e-02	-4.31



(a) $\widehat{F}_1$ with $h =$ 5.0E-02 and $\alpha = 1.5$.



(b) $\widehat{F}_4$ with $h =$ 2.5E-02.

Figure 2.1: Computed solutions of Example 2.1 using $\widehat{F}_1$ and $\widehat{F}_4$ with the standard initial guess.

The schemes with $\widehat{F}_2$ and $\widehat{F}_3$ exhibit behavior similar to that with $\widehat{F}_1$, and $\widehat{F}_5$ exhibits behavior similar to $\widehat{F}_4$. By Lemma 2.2, we consider a quadratic order of convergence to be optimal. Thus, by Table 2.1, the Lax-Friedrichs-like schemes do

exhibit an optimal quadratic order of convergence. On the other hand, the Godunov-like schemes converge inconsistently. This inconsistency is mostly due to *fsolve* failing to find a root.

If we fix our initial guess as the approximation computed by $\widehat{F}_1$ with $\alpha = 1.5$ and $h = 0.1$, we get the results of Table 2.2. Thus, Godunov-like schemes converge with high levels of accuracy when the nonlinear solver has a sufficiently good initial guess. Since *fsolve* is very sensitive towards the initial guess for Godunov-like schemes, it is hard to characterize a rate of convergence. We also observe in Table 2.1 that the error for $h = 0.1, 0.05, 0.025, 0.00625$ is consistent with the error of the initial guess for the Godunov-like schemes. In contrast, the Lax-Friedrichs-like schemes converge for a much wider range of initial guesses when using the nonlinear solver *fsolve*.

Table 2.2: Rates of convergence for Example 2.1 using $\widehat{F}_1$ and $\widehat{F}_4$ with an improved initial guess.

	$\widehat{F}_1$, $\alpha = 1.5$		$\widehat{F}_4$	
h	error	order	error	order
1.0000e-01	2.71e-02		8.24e-08	
5.0000e-02	5.10e-03	2.41	1.58e-06	-4.26
2.5000e-02	1.03e-03	2.31	1.60e-05	-3.34
1.2500e-02	2.33e-04	2.14	9.06e-05	-2.51
6.2500e-03	5.58e-05	2.06	1.42e-02	-7.29

Example 2.2. *This example concerns the 1-D Monge-Ampère equation. Consider the problem*

$$-u_{xx}^2 + 1 = 0, \qquad 0 < x < 1,$$
$$u(0) = 0, \quad u(1) = 1/2.$$

This problem has exactly two solutions

$$u^+(x) = \frac{1}{2}x^2, \qquad u^-(x) = -\frac{1}{2}x^2 + x,$$

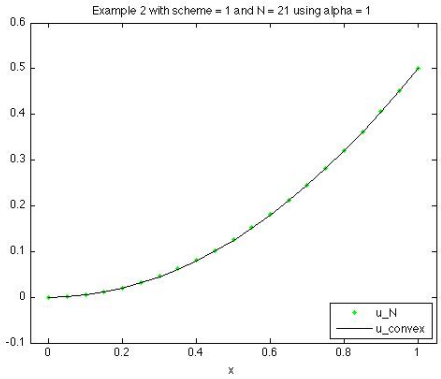
where u^+ is convex and u^- is concave. However, u^+ is the unique viscosity solution that preserves the ellipticity of the operator.

Using $U^{(0)}$ as the linear interpolant of the boundary data, the computed results with both types of schemes are given in Table 2.3 and Figure 2.2. We note that the Lax-Friedrichs-like schemes converge to the unique ellipticity preserving solution (i.e., convex solution) for $\alpha > 0$ sufficiently large. However, if $\alpha < 0$ with $|\alpha|$ sufficiently large, the Lax-Friedrichs-like schemes converge to u^-. We have u^- is the unique viscosity solution for the PDE $u_{xx}^2 - 1 = 0$. Forming the corresponding Lax-Friedrichs-like scheme and multiplying by -1 is equivalent to letting $\alpha < 0$ in the above formulation. Thus, the convergence to u^- for $\alpha < 0$ is a positive result.

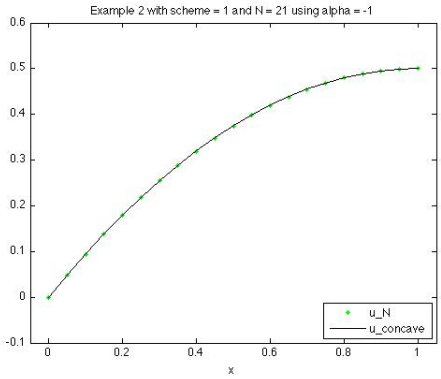
Table 2.3: Rates of convergence for Example 2.2 using $\widehat{F}_1$ with both a positive and negative value for α and $\widehat{F}_4$ all with the standard initial guess.

	$\widehat{F}_1$, $\alpha = 1$		$\widehat{F}_1$, $\alpha = -1$		$\widehat{F}_4$	
h	error	order	error	order	error	order
1.000e-01	2.54e-03		2.54e-03		1.17e-01	
5.000e-02	6.36e-04	2.00	6.36e-04	2.00	1.21e-01	-0.05
2.500e-02	1.59e-04	2.00	1.59e-04	2.00	1.24e-01	-0.04

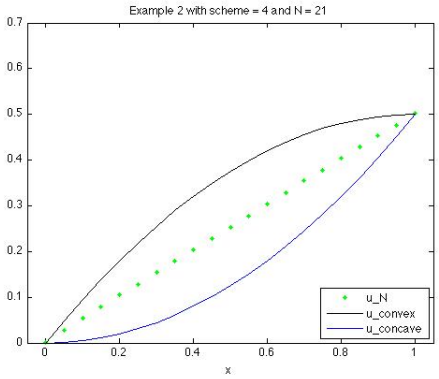
We also test the benefit of using a Lax-Friedrichs-like scheme as opposed to the standard three-point finite difference method. We approximate u using $\widehat{F}_2$ for varying values of α, using the linear interpolant of the boundary data as our initial guess. The computed results are given in Table 2.4 and Figure 2.3. We remark that letting $\alpha = 0$ in $\widehat{F}_2$ corresponds to the standard three-point FD method, which does not converge in the above example. Instead, it often converges to numerical artifacts,



(a) $\widehat{F}_2$ with $h =$ 5.0E-02 and $\alpha = 1$.



(b) $\widehat{F}_2$ with $h =$ 5.0E-02 and $\alpha = -1$.



(c) $\widehat{F}_4$ with $h =$ 5.0E-02.

Figure 2.2: Computed solutions for Example 2.2 using $\widehat{F}_2$ with both a positive and negative value for α and $\widehat{F}_4$ all with the standard initial guess.

that is, solutions of the algebraic system of equations that do not correspond to actual PDE solutions. In contrast, for Godunov-like schemes, the solver does not find an algebraic solution. Thus, the Lax-Friedrichs-like schemes have a mechanism for giving the nonlinear solver a good direction towards finding a root. When α is sufficiently large, the schemes converge. When α is not sufficiently large, while the schemes may not converge, they have a tendency to move towards the correct solution. Furthermore, we can see that the Lax-Friedrichs-like schemes converge quadratically for α bigger than the theoretical lower bound with only a small cost in the level of accuracy. Thus, when dealing with a problem that has an unknown optimal bound for

α, large α values can be used. A shooting method for decreasing α allows the scheme to gain accuracy while maintaining the benefits of the Lax-Friedrichs-like schemes.

Table 2.4: Rates of convergence for Example 2.2 using $\widehat{F}_2$ with decreasing values for α all with the standard initial guess.

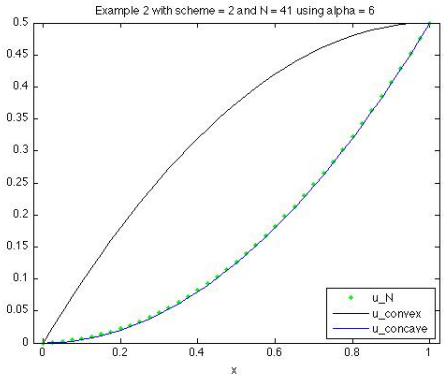
	$\widehat{F}_2$, $\alpha = 6$		$\widehat{F}_2$, $\alpha = 0.05$		$\widehat{F}_2$, $\alpha = 0$	
h	error	order	error	order	error	order
1.000e-01	3.07e-02		1.18e-01		9.00e-02	
5.000e-02	8.51e-03	1.85	3.31e-02	1.83	1.15e-01	-0.35
2.500e-02	2.14e-03	1.99	3.03e-02	0.13	1.15e-01	-0.00

If we first use $\widehat{F}_1$ with $\alpha = 1$ to approximate u on a coarse grid with $h = 0.1$, and then we interpolate the result to get an initial guess for the two proposed schemes and the three-point FD method, we get the results of Table 2.5 and Figure 2.4. Thus, we see that the Godunov-like schemes and the standard FD formulation now converge to u^+ with high levels of accuracy given a sufficiently good initial guess. In fact, they both converge to the same limit.

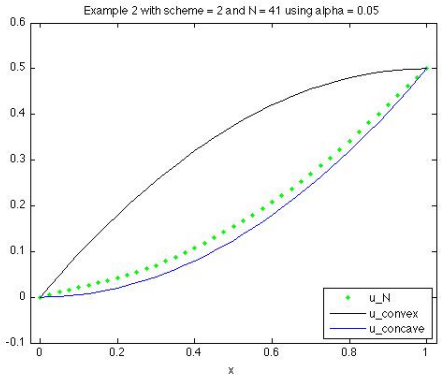
Table 2.5: Rates of convergence for Example 2.2 using $\widehat{F}_1$ with $\alpha = 1$, $\widehat{F}_4$, and the standard three-point scheme all with an improved initial guess.

	$\widehat{F}_1$, $\alpha = 1$		$\widehat{F}_4$		$\widehat{F}_2$, $\alpha = 0$	
h	error	order	error	order	error	order
1.000e-01	2.54e-03		9.96e-15		9.96e-15	
5.000e-02	6.36e-04	2.00	4.54e-13	-5.51	4.54e-13	-5.51
2.500e-02	1.59e-04	2.00	1.46e-10	-8.33	1.46e-10	-8.33
1.250e-02	3.97e-05	2.00	9.85e-10	-2.75	9.85e-10	-2.75

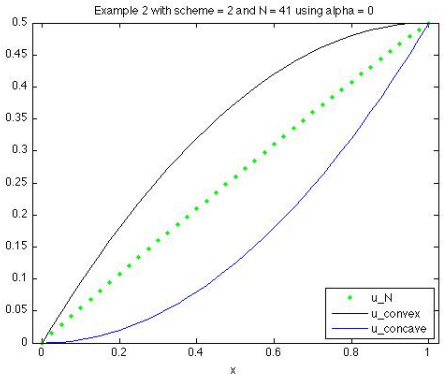
To the contrary, if we use $\widehat{F}_1$ with $\alpha = -1$ to approximate u on a coarse grid with $h = 0.1$ and then interpolate the result as an initial guess, we obtain the results of



(a) $\widehat{F}_2$ with $h =$ 2.5E-02 and $\alpha = 6$.



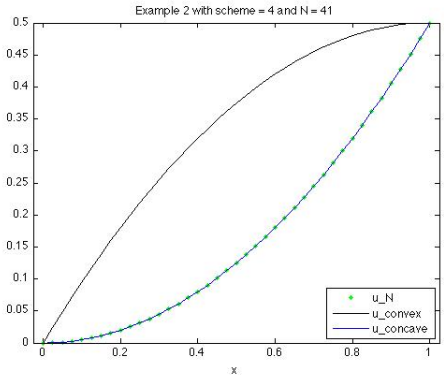
(b) $\widehat{F}_2$ with $h =$ 2.5E-02 and $\alpha = 0.05$.



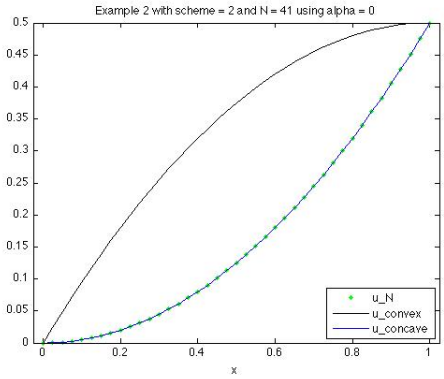
(c) $\widehat{F}_2$ with $h =$ 2.5E-02 and $\alpha = 0$.

Figure 2.3: Computed solutions for Example 2.2 using $\widehat{F}_2$ with decreasing values for α all with the standard initial guess.

Table 2.6 and Figure 2.5. Clearly, none of the schemes converge to u^+. Moreover, the Lax-Friedrichs-like schemes and the Godunov-like schemes do not converge to u^- even if $U^{(0)}$ is close to u^-. Instead, *fsolve* finds no solution when using the two proposed schemes. Thus, the Lax-Friedrichs-like schemes and the Godunov-like schemes appear to only approximate u^+. Since u^+ is the unique viscosity solution of the PDE, lack of convergence to u^- for the Lax-Friedrichs-like schemes for $\alpha > 0$ sufficiently large and for the Godunov-like schemes is strong evidence that our proposed methods have a built-in mechanism to allow for selectivity when approximating viscosity solutions of fully nonlinear PDEs. The standard three-point FD method does converge to u^-.



(a) $\widehat{F}_4$ with $h =$ 2.5E-02 and $\alpha = 6$.



(b) $\widehat{F}_2$ with $h =$ 2.5E-02 and $\alpha = 0.05$.

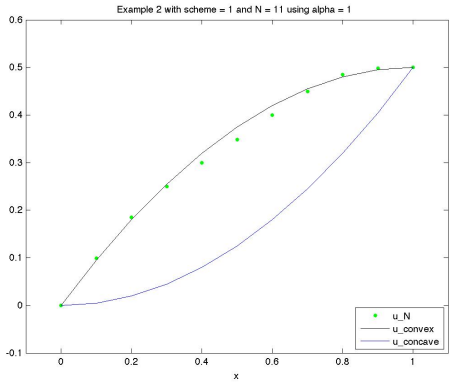
Figure 2.4: Computed solutions for Example 2.2 using $\widehat{F}_4$ and the standard three-point scheme all with an improved initial guess.

When given a sufficiently good guess, the three-point finite difference method will converge to any one of the two solutions. Furthermore, the discretization can create artificial numerical solutions (i.e., numerical artifacts). On the other hand, the g-monotonicity of our proposed schemes appears to prevent having multiple solutions and numerical artifacts.

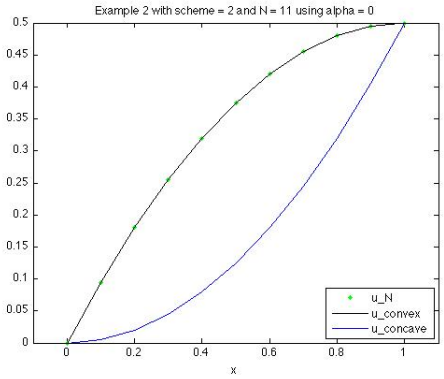
Table 2.6: Rates of convergence for Example 2.2 using $\widehat{F}_1$ with $\alpha = -1$, $\widehat{F}_4$, and the standard three-point scheme all with a poor initial guess.

	$\widehat{F}_1$, $\alpha = -1$		$\widehat{F}_4$		$\widehat{F}_2$, $\alpha = 0$	
h	error	order	error	order	error	order
1.000e-01	2.68e-02		2.56e-03		2.24e-14	
5.000e-02	5.61e-03	2.25	2.54e-03	0.01	8.82e-13	-5.30
2.500e-02	1.26e-02	-1.16	2.54e-03	0.00	8.83e-12	-3.32
1.250e-02	1.41e-02	-0.17	2.54e-03	-0.00	1.63e-09	-7.53

The next two examples deal with Bellman type equations.



(a) $\widehat{F}_1$ with $h =$ 1.0E-01 and $\alpha = 1$.



(b) $\widehat{F}_2$ with $h =$ 1.0E-01 and $\alpha = 0$.

Figure 2.5: Computed solutions for Example 2.2 using $\widehat{F}_1$ with $\alpha = 1$ and the standard three-point scheme all with a poor initial guess.

Example 2.3. *Consider the problem*

$$\min_{\theta(x)\in\{1,2\}} \left\{-A_\theta u_{xx} - S\left(x\right)\right\} = 0, \qquad -1 < x < 1,$$
$$u(-1) = -1, \quad u(1) = 1.$$

for

$$A_1 = 1, \quad A_2 = 2, \quad S(x) = \begin{cases} 12x^2, & \text{if } x < 0, \\ -24x^2, & \text{if } x \geq 0. \end{cases}$$

This problem has the exact solution $u(x) = x\left|x\right|^3$. *We also note that this problem involves an optimization over a finite dimensional set.*

Using the linear interpolant as the initial guess, we obtain the results of Table 2.7 and Figure 2.6. We observe that the Godunov-like scheme converges for larger h values when a root is found, and both schemes exhibit quadratic convergence for this example.

Table 2.7: Rates of convergence for Example 2.3 using $\widehat{F}_1$ with $\alpha = 1$ and $\widehat{F}_4$ all with the standard initial guess.

h	$\widehat{F}_1$, $\alpha = 1$		$\widehat{F}_4$	
	error	order	error	order
1.000e-01	1.29e-01		9.60e-03	
5.000e-02	4.67e-02	1.46	2.50e-03	1.94
2.500e-02	1.46e-02	1.68	6.25e-04	2.00
1.250e-02	4.18e-03	1.80	4.70e-01	-9.55
6.250e-03	1.13e-03	1.89	4.72e-01	-0.01
3.125e-03	2.95e-04	1.93	4.72e-01	-0.00

Example 2.4. *Let $\theta : \mathbb{R} \to \mathbb{R}$ such that $\theta \in L^\infty([2,4])$, and consider the problem*

$$\inf_{-1 \le \theta(x) \le 1} \left\{ -\theta\, u_{xx} + \theta^2\, u + x^{-2} \right\} = 0, \qquad 2 < x < 4,$$

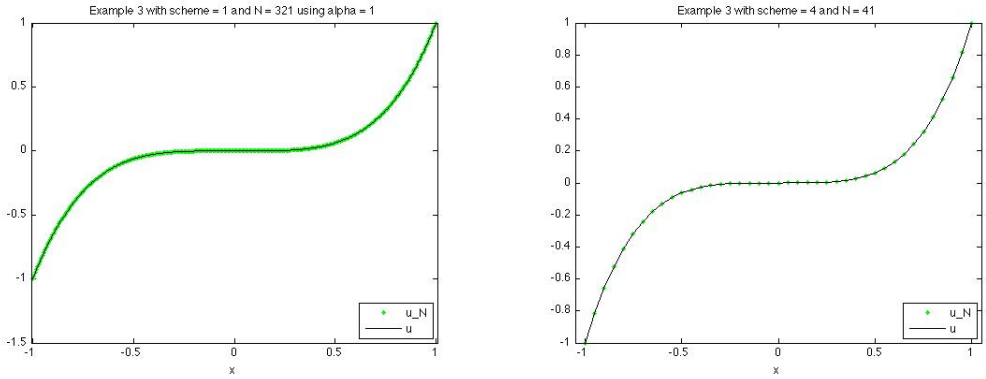
$$u(2) = 4, \quad u(4) = 16.$$

This problem has the exact solution $u(x) = x^2$ which corresponds to $\theta(x) = x^{-2}$.

Let the initial guess be given by the linear interpolant of the boundary data. Then, we obtain the results of Table 2.8. Both schemes have a hard time finding a root for h small, although the Lax-Friedrichs-like schemes do converge towards u for larger values of h.

Table 2.8: Rates of convergence for Example 2.4 using $\widehat{F}_1$ with $\alpha = 0.5$ and $\widehat{F}_4$ all with the standard initial guess.

h	$\widehat{F}_1$, $\alpha = 0.5$		$\widehat{F}_4$	
	error	order	error	order
1.000e-01	3.07e-01		5.59e-01	
5.000e-02	9.88e-02	1.64	4.96e-01	0.17
2.500e-02	3.09e-02	1.68	5.10e+00	-3.36



(a) $\widehat{F}_1$ with $h =$ 6.25E-03 and $\alpha = 1$. **(b)** $\widehat{F}_4$ with $h =$ 5.0E-02.

Figure 2.6: Computed solutions for Example 2.3 using $\widehat{F}_1$ with $\alpha = 1$ and $\widehat{F}_4$ all with the standard initial guess.

Now we choose the initial guess

$$U^{(0)} = \frac{3}{14}x^3 + \frac{16}{7},$$

a simple cubic function that satisfies the boundary conditions. Then, $\|U^{(0)} - u\|_{L^\infty([2,4])} \approx 0.94$, and we get the results of Table 2.9 and Figure 2.7. Thus, the Lax-Friedrichs-like schemes again converge with a rate of almost 2. Also, the Godunov-like schemes converge with high levels of accuracy for $h \geq 0.0125$, but for smaller h, *fsolve* fails to find a root.

We remark that this problem can also be approximated by using a splitting algorithm. The operator can be split into an optimization problem for θ and a linear PDE problem for u, and then a natural scheme is to successively approximate θ and u starting with an initial guess for θ (see Section 5.1). For the above approximations, the nonlinearity due to the infimum was preserved inside the definition of the operator.

Table 2.9: Rates of convergence for Example 2.4 using $\widehat{F}_1$ with $\alpha = 0.5$ and $\widehat{F}_4$ all with an improved initial guess.

h	$\widehat{F}_1$, $\alpha = 0.5$		$\widehat{F}_4$	
	error	order	error	order
1.000e-01	3.07e-01		6.74e-10	
5.000e-02	9.88e-02	1.64	7.04e-08	-6.71
2.500e-02	3.09e-02	1.68	3.41e-09	4.37
1.250e-02	9.02e-03	1.78	8.09e-08	-4.57
6.250e-03	2.47e-03	1.87	9.44e-01	-23.48

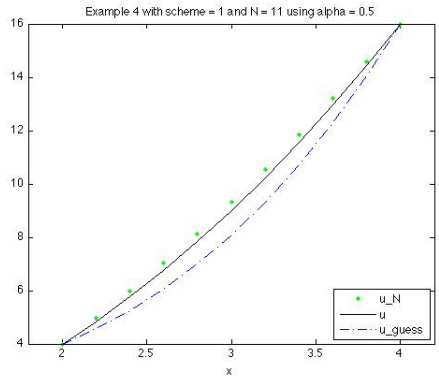
Example 2.5. *Consider the problem*

$$-u_{xx}^3 + 8\ \text{sign}(x) = 0, \qquad -1 < x < 1,$$
$$u(-1) = -1, \quad u(1) = 1,$$

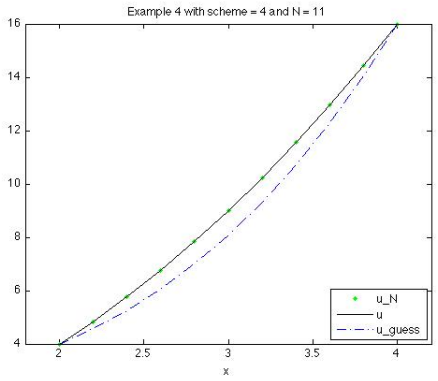
with the exact solution $u(x) = x|x| \in C^1([-1, 1])$. *Note, this example does not have a classical solution.*

Using the linear interpolant of the boundary data as the initial guess, we obtain the results of Table 2.10 and Figure 2.8. We clearly see the quadratic rate of convergence for the Lax-Friedrichs-like schemes. The Godunov-like schemes only converge for $h = 0.0125$. For larger h, the scheme returns the initial guess after failing to find a root. For the test with smaller h, the scheme returns a slightly improved approximation after reaching the maximum number of iterations.

If we fix our initial guess as the approximation formed by $\widehat{F}_1$ with $\alpha = 1.5$ and $h = 0.1$, we then get the results of Table 2.11. As observed in the previous examples, we see that the Godunov-like schemes converge quickly with high levels of accuracy, thus making it difficult to characterize a general rate of convergence.



(a) $\widehat{F}_1$ with $h =$ 2.0E-01 and $\alpha = 0.5$.

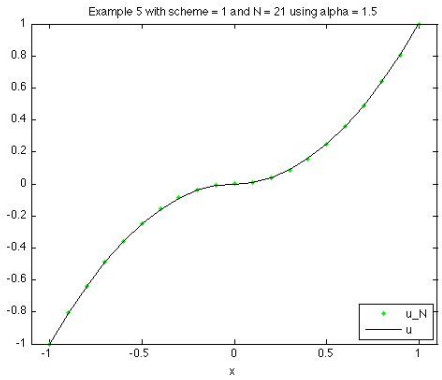


(b) $\widehat{F}_4$ with $h =$ 2.0E-01.

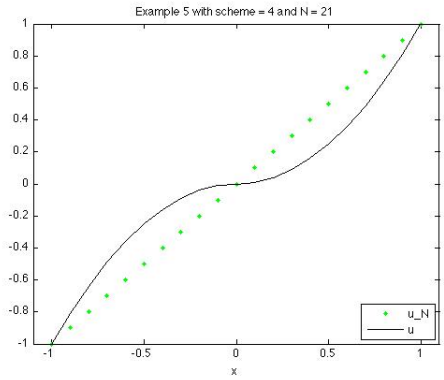
Figure 2.7: Computed solutions for Example 2.4 using $\widehat{F}_1$ with $\alpha = 0.5$ and $\widehat{F}_4$ all with an improved initial guess.

Table 2.10: Rates of convergence for Example 2.5 using $\widehat{F}_1$ with $\alpha = 1.5$ and $\widehat{F}_4$ all with the standard initial guess.

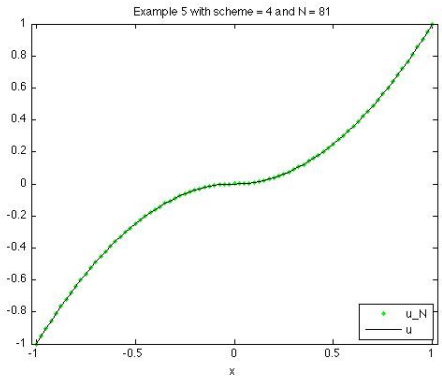
	$\widehat{F}_1$, $\alpha = 1.5$		$\widehat{F}_4$	
h	error	order	error	order
1.000e-01	1.59e-02		2.40e-01	
5.000e-02	3.76e-03	2.08	2.50e-01	-0.06
2.500e-02	9.40e-04	2.00	2.50e-01	0.00
1.250e-02	2.35e-04	2.00	6.69e-06	15.19
6.250e-03	5.88e-05	2.00	2.05e-01	-14.90



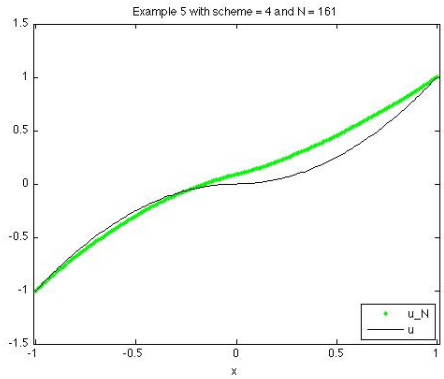
(a) $\widehat{F}_1$ with $h =$ 1.0E-01 and $\alpha = 1.5$.



(b) $\widehat{F}_4$ with $h =$ 1.0E-01.



(c) $\widehat{F}_4$ with $h =$ 2.5E-02.



(d) $\widehat{F}_4$ with $h =$ 1.25E-02.

Figure 2.8: Computed solutions for Example 2.5 using $\widehat{F}_1$ with $\alpha = 1.5$ and $\widehat{F}_4$ all with the standard initial guess.

Table 2.11: Rates of convergence for Example 2.5 using $\widehat{F}_1$ with $\alpha = 1.5$ and $\widehat{F}_4$ all with an improved initial guess.

	$\widehat{F}_1$, $\alpha = 1.5$		$\widehat{F}_4$	
h	error	order	error	order
1.000e-01	1.59e-02		1.84e-08	
5.000e-02	3.76e-03	2.08	4.05e-06	-7.78
2.500e-02	9.40e-04	2.00	8.85e-06	-1.13
1.250e-02	2.35e-04	2.00	6.50e-06	0.45
6.250e-03	5.88e-05	2.00	7.78e-06	-0.26

2.4 Extensions of the New FD Framework to Second Order Parabolic Problems in One Spacial Dimension

In this section we generalize the new FD framework for approximating the viscosity solution $u \in \mathcal{A} \subset C^0\big((a,b)\times(0,T]\big)$ for the second order parabolic problem

$$u_t + F\left(D^2u, \nabla u, u, x, t\right) = 0, \qquad (x,t) \in (a,b)\times(0,T], \tag{2.57a}$$

$$u(a,t) = u_a(t), \qquad t \in [0,T], \tag{2.57b}$$

$$u(b,t) = u_b(t), \qquad t \in [0,T], \tag{2.57c}$$

$$u(x,0) = u_0(x), \qquad x \in (a,b), \tag{2.57d}$$

where the operator F is a continuous, possibly nonlinear, elliptic operator and $\mathcal{A}$ is a function class in which the viscosity solution u is unique. The idea will be to develop fully discrete approximations by using the above FD methodology for elliptic problems to discretize the spatial variable and applying the method of lines. We will consider both implicit and explicit methods for the time-discretization.

2.4.1 Formulation of the Fully Discrete Framework

We first discretize the spatial variable using the same methodology as above. Let $\mathcal{T}_h$ be a grid for $\overline{\Omega}$. Assume U is a grid function defined on $\mathcal{T}_h \times [0,T]$. Substituting a numerical operator $\widehat{F}$ for F in (2.57a), we obtain the semi-discrete equation

$$\frac{\partial}{\partial t}U_i + \widehat{F}\left(\delta^2_{x,h}U_{i-1}, \delta^2_{x,h}U_i, \delta^2_{x,h}U_{i+1}, \delta^+_{x,h}U_i, \delta^-_{x,h}U_i, U_i, x_i, t\right) = 0 \tag{2.58}$$

for $i = 2, 3, \ldots, J-1$, where $U_1(t) = u_a(t)$, $U_J(t) = u_b(t)$, and $U_0(t)$, $U_{J+1}(t)$ are ghost values for all $t \in [0,T]$. To simplify the presentation, we let $U = (U_2, U_3, \ldots, U_{J-1})$

and

$$\widehat{F}_i[U,t] := \widehat{F}\left(\delta^2_{x,h}U_{i-1}, \delta^2_{x,h}U_i, \delta^2_{x,h}U_{i+1}, \delta^+_{x,h}U_i, \delta^-_{x,h}U_i, U_i, x_i, t\right).$$

Then, we can rewrite the semi-discrete equation (2.58) as

$$\frac{\partial}{\partial t}U_i = -\widehat{F}_i[U,t] \tag{2.59}$$

for $i = 2, 3, \ldots, J-1$. Thus, we have a system of $J-2$ first order ordinary differential equations (ODEs) in time.

Next, we discretize the time variable. Pick an integer $N > 0$ and let $\Delta t = T/N$. We use a super-index to denote the approximation at a given time level. Thus, we have U^n_j denotes an approximation for $u(x_j, n\Delta t)$ for $n = 0, 1, \ldots, N$. To incorporate the boundary conditions, (2.57b) and (2.57c), we define $U^n_1 := u_a(n\Delta t)$ and $U^n_J := u_b(n\Delta t)$ for each $n = 0, 1, \ldots, N$. Furthermore, we use the convention that U^n_0 and U^n_{J+1} are ghost values that are dictated by the spatial discretization at each time step $n = 0, 1, \ldots, N$. Lastly, we incorporate the initial condition, (2.57d), by defining $U^0_i := u_0(x_i)$ for each $i = 2, 3, \ldots, J-1$.

Using the above conventions, we can define fully discrete methods for approximating problem (2.57) based on approximating (2.59) using the forward Euler method, the backward Euler method, and the trapezoidal method. Hence, we have the following fully discrete schemes for approximating (2.57):

$$U^{n+1}_i = U^n_i - \Delta t\, \widehat{F}_i\big[U^n, n\Delta t\big], \tag{2.60}$$

$$U^{n+1}_i + \Delta t\, \widehat{F}_i\big[U^{n+1}, (n+1)\Delta t\big] = U^n_i, \tag{2.61}$$

and

$$U^{n+1}_i + \frac{\Delta t}{2}\, \widehat{F}_i\big[U^{n+1}, (n+1)\Delta t\big] = U^n_i - \frac{\Delta t}{2}\, \widehat{F}_i\big[U^n, n\Delta t\big], \tag{2.62}$$

for $i = 2, 3, \ldots, J-1$, $n = 0, 1, \ldots, N-1$, where (2.60), (2.61), and (2.62) correspond to the forward Euler method, backward Euler method, and trapezoidal method, respectively.

Remark 2.7. *Using an implicit method such as the backward Euler method or the trapezoidal method is equivalent to approximating a fully nonlinear elliptic PDE at each time level $n = 1, 2, \ldots, N$. Due to the initial condition, the nonlinear solver has a natural initial guess for each time-step.*

We now consider using Runge-Kutta methods for approximating the system of ODEs given by (2.59). Let ν be a positive integer, $A \in \mathbb{R}^{\nu \times \nu}$, and $b, c \in \mathbb{R}^{\nu}$ such that

$$\sum_{\ell=1}^{\nu} a_{k,\ell} = c_k$$

for each $k = 1, 2, \ldots, \nu$. Pick $k \in \mathbb{R}$ such that $0 \leq k \leq N$, and let V denote a grid function defined on $\mathcal{T}_h$. We define the discrete operator $\widehat{F}_i^k$ by

$$\widehat{F}_i^k[V] := \widehat{F}\left(\delta_{x,h}^2 V_{i-1}, \delta_{x,h}^2 V_i, \delta_{x,h}^2 V_{i+1}, \delta_{x,h}^+ V_i, \delta_{x,h}^- V_i, V_i, x_i, k\,\Delta t\right)$$

for $i = 2, 3, \ldots, J-1$, with

$$V_0 := u_a(k\,\Delta t), \qquad V_J := u_b(k\,\Delta t)$$

and V_1, V_{J+1} two ghost values dictated by the spatial discretization. Then, a generic ν stage Runge-Kutta scheme for approximating (2.59) can be written

$$U_i^{n+1} = U_i^n - \Delta t \sum_{\ell=1}^{\nu} b_\ell\, \widehat{F}_i^{n+c_\ell}\left[\xi^{n,\ell}\right] \tag{2.63}$$

for

$$\xi_i^{n,\ell} = U_i^n - \Delta t \sum_{k=1}^{\nu} a_{k,\ell}\, \widehat{F}_i^{n+c_k}\left[\xi^{n,k}\right]$$

for all $i = 2, 3, \ldots, J-1$, $n = 0, 1, \ldots, N-1$. We note that (2.63) corresponds to an explicit scheme when A is strictly lower diagonal and an implicit scheme otherwise.

Remark 2.8.

(a) We can interpret $\xi_i^{n,\ell}$ in (2.63) as an approximation for $u\big(x_i, (n+c_\ell)\Delta t\big)$. Thus, we use $\widehat{F}_i^{n+c_\ell}$ to enforce the boundary condition at time $t = (n+c_\ell)\Delta t$ by setting $\xi_1^{n,\ell} = u_a\big((n+c_\ell)\Delta t\big)$ and $\xi_J^{n,\ell} = u_b\big((n+c_\ell)\Delta t\big)$ in the definition of $\widehat{F}_i^{n+c_\ell}$.

(b) In practice, the time-discretization method should be chosen to match the order of the spatial discretization. Suppose the time-discretization method has order r. Then, the time-step size Δt should be chosen such that $\Delta t = h^{2/r}$ when using implicit methods without a CFL condition.

2.4.2 Numerical Experiments

We now implement the proposed methodology for a series of parabolic problems. We will test both the implicit trapezoidal method and the explicit midpoint method, a two-stage Runge-Kutta method that corresponds to the tableau in Figure 2.9. A uniform partition for the spatial and time coordinates will be used. For implicit methods, the Matlab built-in nonlinear solver *fsolve* will be used, and the initial guess will be the grid function U from the previous time step, where the initial grid approximation will be formed by the initial condition (2.57d). For the implicit tests, we record the time-step size Δt, and for the explicit tests we record a scaling parameter κ_t, where the time-step size will be given by the assumed CFL condition $\Delta t = \kappa_t h^2$. All errors will be measure in the ℓ^∞ norm defined on the specified grid, and all tests will use the numerical operator $\widehat{F}_2$ defined in (2.40b). We will observe that the proposed scheme appears to converge quadratically when choosing $\beta = 0$ and linearly when choosing $\beta > 0$. Also, we note that the explicit schemes appeared unstable for $\kappa_t \geq 0.01$.

$$\begin{array}{c|c} c & A \\ \hline & b^t \end{array} \quad = \quad \begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array}$$

Figure 2.9: The tableau for the midpoint method.

Example 2.6. *Consider the problem*

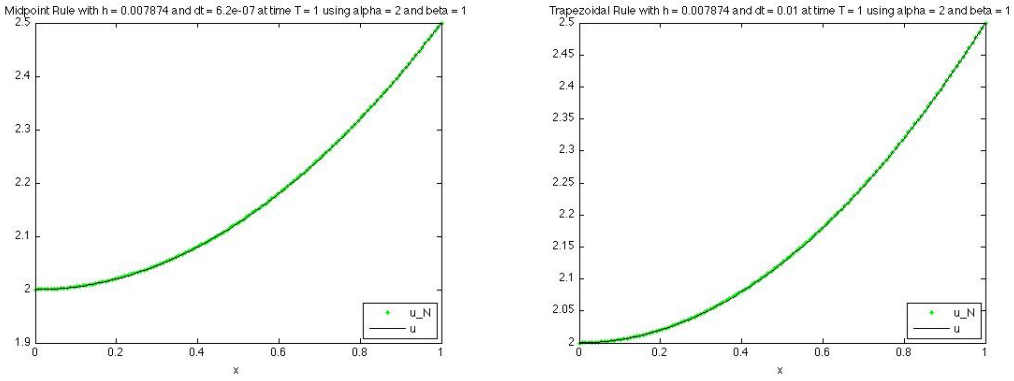
$$\begin{aligned} u_t - u_{xx}\, u &= f \qquad && \text{in } \Omega \times (0,1], \\ u &= g \qquad && \text{on } \partial\Omega \times (0,1], \\ u &= u_0 \qquad && \text{in } \Omega \times \{0\}, \end{aligned}$$

where $\Omega = (0,1)$, $f(x,t) = -\frac{1}{2}x^2 - t^4 + 4\,t^3 - 1$, *and* g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = 0.5\,x^2 + t^4 + 1$.

The problem is actually quasi-linear, with a product nonlinearity. The numerical results are recorded in Table 2.12 and Figure 2.10, where we observe quadratic rates when $\beta = 0$ and better than linear rates when $\beta = 1$.

Table 2.12: Rates of convergence in space for Example 2.6 at time $t = 1.0$ using $\alpha = 2$ and $U_i^0 = u_0(x_i)$.

	Midpoint with $\kappa_t = 0.001$				Trapezoidal with $\Delta t = 0.01$			
	$\beta = 1$		$\beta = 0$		$\beta = 1$		$\beta = 0$	
h	Error	Order	Error	Order	Error	Order	Error	Order
1.43e-01	2.58e-02		1.99e-02		2.58e-02		1.99e-02	
6.67e-02	8.80e-03	1.41	4.84e-03	1.85	8.81e-03	1.41	4.86e-03	1.85
3.23e-02	3.25e-03	1.37	1.15e-03	1.98	3.26e-03	1.37	1.16e-03	1.97
1.59e-02	1.34e-03	1.25	2.78e-04	2.00	1.36e-03	1.24	2.90e-04	1.95
7.87e-03	6.03e-04	1.14	6.83e-05	2.00	6.16e-04	1.13	8.08e-05	1.82



(a) Midpoint with $\kappa_t = 0.001$. **(b)** Trapezoidal with $\Delta t = 0.01$.

Figure 2.10: Computed solutions at time $t = 1.0$ for Example 2.6 using $\alpha = 2$, $\beta = 1$, $h =$ 7.87e-03, and $U_i^0 = u_0(x_i)$.

Example 2.7. *Consider the problem*

$$\begin{aligned} u_t - u_x \ln\big(u_{xx} + 1\big) &= f \qquad && \text{in } \Omega \times (0, 1/2], \\ u &= g \qquad && \text{on } \partial\Omega \times (0, 1/2], \\ u &= u_0 \qquad && \text{in } \Omega \times \{0\}, \end{aligned}$$

where $\Omega = (0, 2)$, $f(x,t) = -e^{(t+1)x}\Big(x - (t+1)\ln\big((t+1)^2 e^{(t+1)x} + 1\big)\Big)$, *and* g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = e^{(t+1)x}$.

Notice, the exact solution u cannot be factored into the form $u(x,t) = G(t)\,Y(x)$ for some functions G and Y. The results for the trapezoidal method and the midpoint method are recorded in Table 2.13 and Figure 2.11, where we observe less than quadratic rates of convergence for $\beta = 0$ and better than linear rates of convergence with $\beta = 1$.

Table 2.13: Rates of convergence in space for Example 2.7 at time $t = 0.5$ using $\alpha = 2$ and $U_i^0 = u_0(x_i)$.

	Midpoint with $\kappa_t = 0.005$				Trapezoidal with $\Delta t = 0.005$			
	$\beta = 1$		$\beta = 0$		$\beta = 1$		$\beta = 0$	
h	Error	Order	Error	Order	Error	Order	Error	Order
2.86e-01	1.98e+00		1.57e+00		1.98e+00		1.57e+00	
1.33e-01	8.62e-01	1.09	6.45e-01	1.17	8.58e-01	1.10	6.43e-01	1.17
6.45e-02	3.42e-01	1.27	2.38e-01	1.38	3.39e-01	1.28	2.36e-01	1.38
3.17e-02	1.32e-01	1.34	7.78e-02	1.57	1.29e-01	1.37	7.71e-02	1.58
1.57e-02	5.66e-02	1.21	2.32e-02	1.73	5.15e-02	1.31	2.29e-02	1.73

Example 2.8. *Consider the Hamilton-Jacobi-Bellman problem*

$$
\begin{aligned}
u_t - \min_{\theta(t,x)\in\{1,2\}} \Big\{ A_\theta \, u_{xx} - c(x,t)\cos(t)\,\sin(x) - \sin(t)\,\sin(x) \Big\} = 0 \quad & \text{in } \Omega \times (0, \frac{1}{2}], \\
u = g \quad & \text{on } \partial\Omega \times (0,1], \\
u = u_0 \quad & \text{in } \Omega \times \{0\},
\end{aligned}
$$

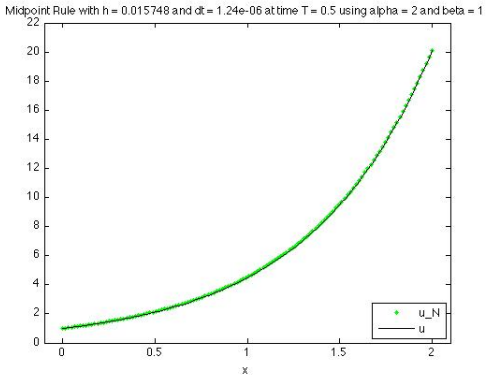
where $\Omega = (0, 2\pi)$, $A_1 = 1, A_2 = \frac{1}{2}$,

$$
c(x,t) = \begin{cases} 1, & \text{if } 0 < t \le \frac{\pi}{2} \text{ and } 0 < x \le \pi \text{ or } \frac{\pi}{2} < t \le \pi \text{ and } \pi < x < 2\pi, \\ \frac{1}{2}, & \text{otherwise,} \end{cases}
$$

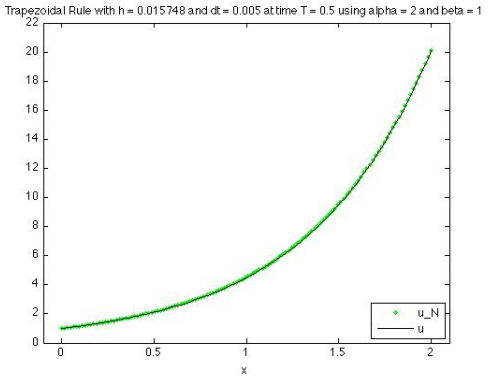
and g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = \cos(t)\sin(x)$.

Notice that this problem involves an optimization over a finite dimensional set, and the solution corresponds to

$$
\theta(x,t) = \begin{cases} 1, & \text{if } c(x,t) = 1, \\ 2, & \text{if } c(x,t) = \frac{1}{2}. \end{cases}
$$



(a) Midpoint with $\kappa_t = 0.005$.



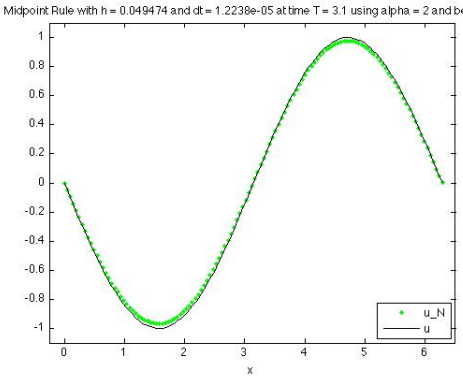
(b) Trapezoidal with $\Delta t = 0.005$.

Figure 2.11: Computed solutions at time $t = 0.5$ for Example 2.7 using $\alpha = 2$, $\beta = 1$, $h =$ 1.57e-02, and $U_i^0 = u_0(x_i)$.

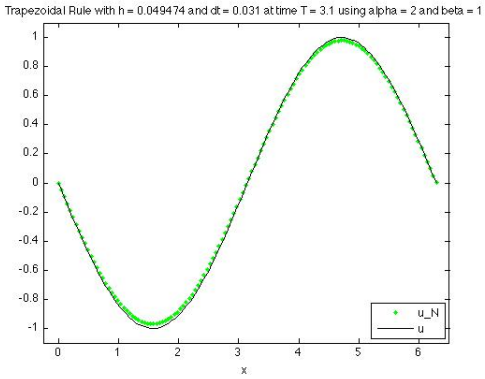
The results are recorded in Table 2.14 and Figure 2.12, where we again observe quadratic rates of convergence for $\beta = 0$ and sub-quadratic rates of convergence for $\beta = 1$.

Table 2.14: Rates of convergence in space for Example 2.8 at time $t = 3.1$ using $\alpha = 2$ and $U_i^0 = u_0(x_i)$.

	Midpoint with $\kappa_t = 0.005$				Trapezoidal with $\Delta t = 0.031$			
	$\beta = 1$		$\beta = 0$		$\beta = 1$		$\beta = 0$	
h	Error	Order	Error	Order	Error	Order	Error	Order
8.98e-01	7.41e-01		5.99e-01		7.42e-01		6.00e-01	
4.19e-01	3.88e-01	0.85	1.93e-01	1.49	3.89e-01	0.85	1.93e-01	1.49
2.03e-01	1.63e-01	1.20	4.71e-02	1.94	1.63e-01	1.20	4.65e-02	1.96
9.97e-02	6.96e-02	1.20	1.14e-02	2.00	6.91e-02	1.21	1.04e-02	2.11
4.95e-02	3.17e-02	1.12	2.81e-03	2.00	3.08e-02	1.15	2.37e-03	2.11



(a) Midpoint with $\kappa_t = 0.005$.



(b) Trapezoidal with $\Delta t = 0.031$.

Figure 2.12: Computed solutions at time $t = 3.1$ for Example 2.8 using $\alpha = 2$, $\beta = 1$, $h =$ 4.95e-02, and $U_i^0 = u_0(x_i)$.

Example 2.9. *Consider the Hamilton-Jacobi-Bellman problem*

$$\begin{aligned} u_t - \inf_{-1\le\theta(t,x)\le 1}\Big\{|x-1|u_{xx} + \theta u_x\Big\} &= f \qquad \text{in } \Omega\times(0,\frac{1}{2}],\\ u &= g \qquad \text{on } \partial\Omega\times(0,1],\\ u &= u_0 \qquad \text{in } \Omega\times\{0\}, \end{aligned}$$

where $\Omega = (0,3)$, $f(x,t) = -|x-1|^2\,(|x-1|+3)\,e^{-t}$, *and* g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = |x-1|^3\,e^{-t} \in C^2(0,3)$.

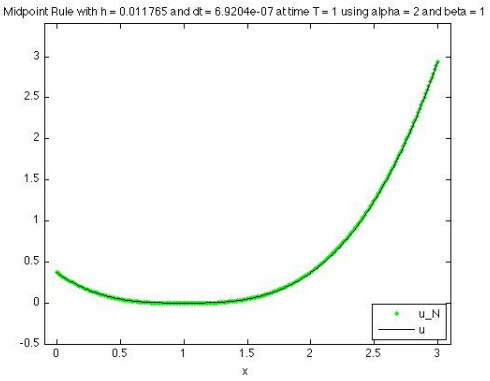
Notice that the above operator is not second order when $x = 1$. Also, the viscosity solution corresponds to

$$\theta(x,t) = \begin{cases} 1, & \text{if } x < 1,\\ -1, & \text{if } x > 1. \end{cases}$$

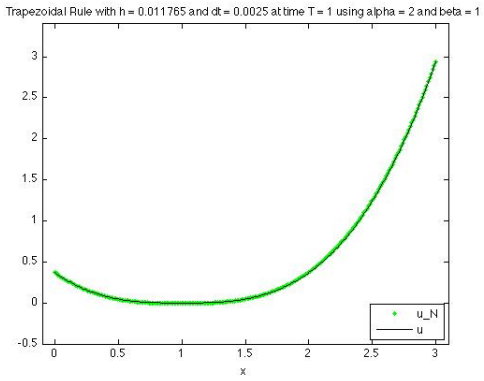
The approximation results are recorded in Table 2.15 and Figure 2.13. Observe, the rates of convergence appear to suffer possibly due to the lower regularity of the viscosity solution.

Table 2.15: Rates of convergence in space for Example 2.9 at time $t = 1.0$ using $\alpha = 2$ and $U_i^0 = u_0(x_i)$.

	Midpoint with $\kappa_t = 0.005$				Trapezoidal with $\Delta t = 0.0025$			
	$\beta = 1$		$\beta = 0$		$\beta = 1$		$\beta = 0$	
h	Error	Order	Error	Order	Error	Order	Error	Order
2.00e-01	2.20e-01		4.02e-01		2.20e-01		4.02e-01	
9.68e-02	7.67e-02	1.45	2.05e-01	0.93	7.69e-02	1.45	2.05e-01	0.93
4.76e-02	4.07e-02	0.89	8.11e-02	1.31	4.06e-02	0.90	8.10e-02	1.31
2.36e-02	1.82e-02	1.15	3.09e-02	1.38	1.84e-02	1.13	3.08e-02	1.38
1.18e-02	9.52e-03	0.93	1.20e-02	1.35	9.65e-03	0.92	1.20e-02	1.35



(a) Midpoint with $\kappa_t = 0.005$.



(b) Trapezoidal with $\Delta t = 0.0025$.

Figure 2.13: Computed solutions at time $t = 1.0$ for Example 2.9 using $\alpha = 2$, $\beta = 1$, $h =$ 1.18e-02, and $U_i^0 = u_0(x_i)$.

2.5 Extensions of the New FD Framework to Second Order Elliptic Problems in Higher Dimensions

We now generalize the new FD framework for second order elliptic problems in high-dimensions. The analysis for the proposed method is still open. As such, the discussion about the analysis of the high-dimensional framework will be postponed until Section 6.1. The proposed discretizations will be tested in Section 2.4.2.

2.5.1 Formulation of the Framework

We present the most natural generalization of the one-dimensional framework, given by: Find a grid function U such that

$$\widehat{F}\left(D_h^{++}U_\alpha, D_h^{+-}U_\alpha, D_h^{-+}U_\alpha, D_h^{--}U_\alpha, \nabla_h^{+}U_\alpha, \nabla_h^{-}U_\alpha, U_\alpha, x_\alpha\right) = 0 \tag{2.64}$$

for all $\alpha \in \mathbb{N}_J^d$ such that $\alpha_i \in \{2, 3, \ldots, J_i - 1\}$ for all $i = 1, 2, \ldots, d$. As expected, U_α is intended to be an approximation of $u(x_\alpha)$ for all $x_\alpha \in \mathcal{T}_h$, and U_α is a ghost value for all α such that $\alpha_i \in \{0, J_i + 1\}$ for some $i \in \{1, 2, \ldots, d\}$.

We have the following natural generalizations of Definitions 2.2 and 2.3:

Definition 2.4. *The function* $\widehat{F} : \left(\mathbb{R}^{d\times d}\right)^4 \times \left(\mathbb{R}^d\right)^2 \times \mathbb{R} \times \Omega \to \mathbb{R}$ *in* (2.64) *is called a* numerical operator. *FD method* (2.64) *is said to be an* admissible *scheme for problem* (2.1) *if it has at least one (grid function) solution* U *such that* $U_\alpha = g(x_\alpha)$ *for all* $x_\alpha \in \partial\Omega$.

Definition 2.5.

(i) *Let* $P \in \overline{\mathbb{R}}^{d\times d}$, $q \in \overline{\mathbb{R}}^{d}$, $v \in \mathbb{R}$, *and* $x \in \overline{\Omega}$. *FD method* (2.64) *is said to be a* consistent *scheme if* $\widehat{F}$ *satisfies*

$$\liminf_{\substack{P^{\mu\,\nu}\to P, q^{\mu}\to q;\mu,\nu\in\{+,-\}\\ \nu\to v,\ \xi\to x}} \widehat{F}(P^{++},P^{+-},P^{-+},P^{--},q^{+},q^{-},\nu,\xi) \geq F_{*}(P,q,v,x),$$

$$\limsup_{\substack{P^{\mu\,\nu}\to P, q^{\mu}\to q;\mu,\nu\in\{+,-\}\\ \nu\to v,\ \xi\to x}} \widehat{F}(P^{++},P^{+-},P^{-+},P^{--},q^{+},q^{-},\nu,\xi) \leq F^{*}(P,q,v,x),$$

where F_{*} *and* F^{*} *denote, respectively, the lower and the upper semi-continuous envelopes of* F. *Thus, we have*

$$F_{*}(P,q,v,x) := \liminf_{\substack{\widetilde{P}\to P,\widetilde{q}\to q,\\ \widetilde{v}\to v,\widetilde{x}\to x}} F\left(\widetilde{P},\widetilde{q},\widetilde{v},\widetilde{x}\right),$$

$$F^{*}(P,q,v,x) := \limsup_{\substack{\widetilde{P}\to P,\widetilde{q}\to q,\\ \widetilde{v}\to v,\widetilde{x}\to x}} F\left(\widetilde{P},\widetilde{q},\widetilde{v},\widetilde{x}\right),$$

where $\widetilde{P} \in \mathbb{R}^{d\times d}$, $\widetilde{q} \in \mathbb{R}^{d}$, $\widetilde{v} \in \mathbb{R}$, *and* $\widetilde{x} \in \Omega$.

(ii) *FD method* (2.64) *is said to be a* generalized monotone (g-monotone) *scheme if for each* $\alpha \in \mathbb{N}_{J}^{d}$ *such that* $x_{\alpha} \in \Omega$, $\widehat{F}(P^{++},P^{+-},P^{-+},P^{--},q^{+},q^{-},v,x_{\alpha})$ *is monotone increasing in* P^{++}, P^{--}, *and* q^{-} *and monotone decreasing in* P^{+-}, P^{-+}, *and* q^{+}. *More precisely, for all* $P^{\mu\,\nu} \in \mathbb{R}^{d\times d}$ *and* $q^{\mu} \in \mathbb{R}^{d}$, $\mu,\nu \in \{+,-\}$, *for all* $v \in \mathbb{R}$, *and for all* $x_{\alpha} \in \Omega$, *there holds*

$$\widehat{F}\left(A,P^{+-},P^{-+},P^{--},q^{+},q^{-},v,x_{\alpha}\right) \leq \widehat{F}\left(B,P^{+-},P^{-+},P^{--},q^{+},q^{-},v,x_{\alpha}\right),$$

$$\widehat{F}\left(P^{++},A,P^{-+},P^{--},q^{+},q^{-},v,x_{\alpha}\right) \geq \widehat{F}\left(P^{++},B,P^{-+},P^{--},q^{+},q^{-},v,x_{\alpha}\right),$$

$$\widehat{F}\left(P^{++},P^{+-},A,P^{--},q^{+},q^{-},v,x_{\alpha}\right) \geq \widehat{F}\left(P^{++},P^{+-},B,P^{--},q^{+},q^{-},v,x_{\alpha}\right),$$

$$\widehat{F}\left(P^{++},P^{+-},P^{-+},A,q^{+},q^{-},v,x_{\alpha}\right) \leq \widehat{F}\left(P^{++},P^{+-},P^{-+},B,q^{+},q^{-},v,x_{\alpha}\right),$$

for all $A, B \in \mathcal{S}^{d\times d}$ *such that* $A \le B$*, where* $A \le B$ *means that* $B - A$ *is a nonnegative definite matrix, and*

$$\widehat{F}\left(P^{++}, P^{+-}, P^{-+}, P^{--}, a, q^-, v, x_\alpha\right) \ge \widehat{F}\left(P^{++}, P^{+-}, P^{-+}, P^{--}, b, q^-, v, x_\alpha\right),$$
$$\widehat{F}\left(P^{++}, P^{+-}, P^{-+}, P^{--}, q^+, a, v, x_\alpha\right) \le \widehat{F}\left(P^{++}, P^{+-}, P^{-+}, P^{--}, q^+, b, v, x_\alpha\right),$$

for all $a, b \in \mathbb{R}^d$ *such that* $a_i \le b_i$ *for all* $i = 1, 2, \ldots, d$*. In other words,* $\widehat{F}(\uparrow, \downarrow, \downarrow, \uparrow, \downarrow, \uparrow, v, x_\alpha)$.

(iii) *Let* (2.64) *be an* admissible *FD method. A solution* U *of* (2.64) *is said to be* stable *if there exists a constant* $C > 0$*, which is independent of* h*, such that* U *satisfies*

$$\|U\|_{\ell^\infty(\mathcal{T}_h)} := \max_{\alpha\in\mathbb{N}_J^d} |U_\alpha| \le C.$$

Also, (2.64) *is said to be a* stable scheme *if all of its solutions are stable solutions.*

Under the generalized high-dimensional framework, the Lax-Friedrichs-like numerical operator takes the form

$$\begin{aligned}\widehat{F}_b\left(P^{++}, P^{+-}, P^{-+}, P^{--}, q^+, q^-, v, x\right) &\\ := F\left(b_1\, P^{++} + b_2\, P^{+-} + b_3\, P^{-+} + b_4\, P^{--}, \frac{q^+ + q^-}{2}, v, x\right) &\\ + \alpha : \left(P^{++} - P^{+-} - P^{-+} + P^{--}\right) - \beta \cdot \left(q^+ - q^-\right), &\end{aligned} \tag{2.66}$$

where $\{b_j\}_{j=1}^4$ are nonnegative constants satisfying $b_1 + b_2 + b_3 + b_4 = 1$, $\alpha \in \mathbb{R}^{d\times d}$ is an undetermined nonnegative matrix or matrix-valued function, and $\beta \in \mathbb{R}^d$ is an undetermined nonnegative vector or vector-valued function.

Remark 2.9.

(a) The term $\alpha : \left(P^{++} - P^{+-} - P^{-+} + P^{--}\right)$ *is called a* numerical moment *due to the fact, for* $i, j \in \{1, 2, \ldots, d\}$,

$$\begin{aligned}\left(\delta^{+}_{x_i,h_i}\delta^{+}_{x_j,h_j} - \delta^{+}_{x_i,h_i}\delta^{-}_{x_j,h_j} - \delta^{-}_{x_i,h_i}\delta^{+}_{x_j,h_j} + \delta^{-}_{x_i,h_i}\delta^{-}_{x_j,h_j}\right)U_\alpha &= h_i h_j \delta^{2}_{x_i,h_i}\delta^{2}_{x_j,h_j}U_\alpha \\ &= h_i h_j \delta^{2}_{x_j,h_j}\delta^{2}_{x_i,h_i}U_\alpha,\end{aligned}$$

an $O\left(h_i^2 + h_j^2\right)$ *approximation of* $u_{x_i x_i x_j x_j}(x_\alpha)$ *scaled by* $h_i h_j$. *Then,*

$$\mathbf{1} : \left(D_h^{++} - D_h^{+-} - D_h^{-+} + D_h^{--}\right) U_\alpha \approx h^2 \Delta^2 u(x_\alpha),$$

where $\mathbf{1}$ *denotes the* $d \times d$ *matrix with all entries equal to* 1.

(b) The above proposed Lax-Friedrichs-like FD method will be tested in Section 2.5.2 for two-dimensional problems with $b_1 = b_4 = 0$ *and* $b_2 = b_3 = \frac{1}{2}$.

Recall that for the convergence proof in one-dimension, we used an additional assumption as stated in part *(d)* of Remark 2.3. Thus, we further assume there exists a function $\widehat{G}$ such that the numerical operator $\widehat{F}$ has the form

$$\widehat{F}\left(P^{++}, P^{+-}, P^{-+}, P^{--}, q^{+}, q^{-}, v, x\right) = \widehat{G}\left(\widetilde{P}, \overline{P}, q^{+}, q^{-}, v, x\right),$$

where

$$\widetilde{P} = \frac{P^{++} + P^{--}}{2}, \qquad \overline{P} = \frac{P^{+-} + P^{-+}}{2}.$$

Furthermore, we assume $\widehat{G}$ is increasing in $\widetilde{P}$ and q^{+}, $\widehat{G}$ is decreasing in $\overline{P}$ and q^{-}, and

$$\liminf_{\substack{\widetilde{P},\overline{P}\to P,\ q^{+},q^{-}\to q \\ \nu\to v,\ \xi\to x}} \widehat{G}(\widetilde{P}, \overline{P}, q^{+}, q^{-}, \nu, \xi) \geq F_*(P, q, v, x), \tag{2.67a}$$

$$\limsup_{\substack{\widetilde{P},\overline{P}\to P,\ q^{+},q^{-}\to q \\ \nu\to v,\ \xi\to x}} \widehat{G}(\widetilde{P}, \overline{P}, q^{+}, q^{-}, \nu, \xi) \leq F^*(P, q, v, x), \tag{2.67b}$$

where F_* and F^* denote, respectively, the lower and the upper semi-continuous envelopes of F.

Using the above assumptions and the discrete Hessians defined by (2.11), the FD method (2.64) becomes: Find a grid function U such that

$$\widehat{G}\left(\widetilde{D}_h^2 U_\alpha, \overline{D}_h^2 U_\alpha, \nabla_h^+ U_\alpha, \nabla_h^- U_\alpha, U_\alpha, x_\alpha\right) = 0 \tag{2.68}$$

for all $\alpha \in \mathbb{N}_J^d$ such that $\alpha_i \in \{2, 3, \ldots, J_i - 1\}$ for all $i = 1, 2, \ldots, d$.

The convergence of the proposed FD method for the general high-dimensional elliptic problem (2.1) remains open. More details on theoretical issues that must be addressed to prove the general convergence will be discussed in Section 6.1. However, we can easily show convergence to the viscosity solution of (2.1) when the differential operator F is a function of only the diagonal entries of the Hessian argument. The proof follows directly from the one-dimensional proof, where we now treat each coordinate direction entirely independent of the others.

To this end, we define the FD method: Find a grid function U such that

$$\widehat{G}\left(\operatorname{diag}\left(\widetilde{D}_h^2 U_\alpha\right), \operatorname{diag}\left(\overline{D}_h^2 U_\alpha\right), \nabla_h^+ U_\alpha, \nabla_h^- U_\alpha, U_\alpha, x_\alpha\right) = 0 \tag{2.69}$$

for all $\alpha \in \mathbb{N}_J^d$ such that $\alpha_i \in \{2, 3, \ldots, J_i - 1\}$ for all $i = 1, 2, \ldots, d$, where, for $A \in \mathbb{R}^{d \times d}$, we have $\operatorname{diag}(A) \in \mathbb{R}^{d \times d}$ defined by

$$[\operatorname{diag}(A)]_{i,j} := \begin{cases} A_{i,i} & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

for all $i, j = 1, 2, \ldots, d$.

Then, we have the following theorem, which states convergence of the piecewise constant extension function defined by

$$u_h(x) := U_\alpha, \qquad \forall x \in \prod_{j=1,2,\ldots,d} (x_\alpha - h_j e_j/2, x_\alpha + h_j e_j/2], \tag{2.70}$$

for a given grid function U, where $\{e_j\}_{j=1}^d$ denotes the canonical basis for $\mathbb{R}^d$.

Theorem 2.5. *Suppose problem (2.1) satisfies the comparison principle of Definition 1.4, has a unique continuous viscosity solution u, and the operator F is independent of $u_{x_i x_j}$ for all $i, j \in \{1, 2, \ldots, d\}$ such that $i \neq j$. Let U be a solution to a consistent, g-monotone, and stable FD method (2.69), and let u_h be its piecewise constant extension given by (2.70). Then u_h converges to u locally uniformly as $h \to 0^+$.*

Remark 2.10. *By Theorem 2.5 and straight-forwardly generalizing the results of Section 2.3.4, we can construct convergent Lax-Friedrichs-like FD methods for the linear second order elliptic Dirichlet boundary value problem*

$$\begin{aligned} -\sum_{i=1}^d a_i u_{x_i x_i} + \sum_{i=1}^d b_i u_{x_i} + c\, u = 0, \qquad & \text{in } \Omega \subset \mathbb{R}^d, \\ u = g, \qquad & \text{on } \partial\Omega, \end{aligned}$$

where the coefficient functions $a_i, b_i, c \geq 0$, $i = 1, 2, \ldots, d$, are all in $L^\infty(\Omega)$.

2.5.2 Numerical Experiments

We now implement and test the FD method defined by (2.64) for a series of nonlinear problems with two spatial dimensions using the Lax-Friedrichs-like numerical operator defined by (2.66) with $b_1 = b_4 = 0$ and $b_2 = b_3 = \frac{1}{2}$. We will again see that the proposed schemes appear to have order two when choosing $\beta = \mathbf{0}$ and order one when choosing $\beta > \mathbf{0}$ for problems with differentiable operators and smooth solutions.

However, for problems with less regular solutions and for Hamilton-Jacobi-Bellman type problems, the rates of convergence appear to suffer accordingly. For all tests, we use the Matlab built in nonlinear solver *fsolve* with an initial guess given by the grid function with all interior nodal values equal to zero and all boundary nodal values given by the Dirichlet boundary condition. We also introduce the auxiliary boundary condition $\Delta u = 0$ by introducing ghost values to enforce the constraint equation

$$\sum_{i=1}^{d} \delta^2_{x_i,h_i} U_\alpha = 0$$

for all $x_\alpha \in \mathcal{T}_h$ such that $x_\alpha \in \partial\Omega$. The additional boundary constraint will be used to define values for $D_h^{\pm\pm} U_\alpha$ for grid points such that $\alpha_i \in \{2, J_i - 1\}$ for some $i \in \{1, 2, \ldots, d\}$. Such a constraint equation is consistent with the observation that the proposed Lax-Friedrichs-like numerical operator yields a direct realization of the vanishing moment method, as seen in Remark 2.9.

Example 2.10. *Consider the Monge-Ampère problem*

$$\begin{aligned} -\det D^2 u = -u_{xx}\, u_{yy} + u_{xy}\, u_{yx} &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega, \end{aligned}$$

where $f = -(1 + x^2 + y^2)e^{x^2+y^2}$, $\Omega = (0,1) \times (0,1)$, *and* g *is chosen such that the viscosity solution is given by* $u(x,y) = e^{\frac{x^2+y^2}{2}}$.

The numerical results are recorded in Table 2.16 and Figure 2.14, where we observe linear rates of convergence when $\beta = \mathbf{0}$ and quadratic rates of convergence when $\beta = 10\mathbf{1}$. We see that even for the trivial initial guess, the proposed scheme with a Newton solver does not converge to a numerical artifact, as discussed in Section 1.3 with respect to the same fully nonlinear boundary value problem using a standard nine-point FD method.

Table 2.16: Rates of convergence for Example 2.10 using $\alpha = 24I$ and *fsolve* with initial guess $U^{(0)} = 0$.

	$\beta = 10\mathbf{1}$		$\beta = \mathbf{0}$	
h	Error	Order	Error	Order
1.29e-01	3.33e-01		4.04e-01	
9.43e-02	2.58e-01	0.83	2.51e-01	1.54
7.44e-02	2.02e-01	1.03	1.59e-01	1.93
6.15e-02	1.64e-01	1.11	1.07e-01	2.08
5.24e-02	1.36e-01	1.16	7.58e-02	2.13

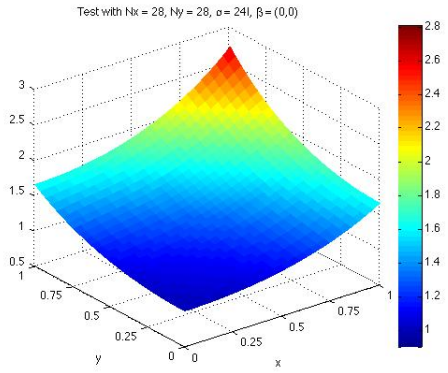


Figure 2.14: Computed solution for Example 2.10 using $\alpha = 24I$, $\beta = \mathbf{0}$, $h =$ 5.24e-02, and *fsolve* with initial guess $U^{(0)} = 0$.

Example 2.11. *Consider the Monge-Ampère problem*

$$\begin{aligned} -\det D^2 u = -u_{xx}\, u_{yy} + u_{xy}\, u_{yx} = 0 \qquad & \text{in } \Omega, \\ u = g \qquad & \text{on } \partial\Omega, \end{aligned}$$

where $\Omega = (-1,1) \times (-1,1)$ *and* g *is chosen such that the viscosity solution is given by* $u(x,y) = |x| \in C^0(\Omega)$.

The numerical results for approximating the given problem can be found in Table 2.17 and Figure 2.15. We observe that the rates of convergence appear to

be linear for both $\beta = \mathbf{1}$ and $\beta = \mathbf{0}$. However, the scheme appears more accurate for $\beta = \mathbf{0}$. The decreased rate of convergence for $\beta = \mathbf{0}$ is expected due to the fact $u \in H^1(\Omega)$.

Table 2.17: Rates of convergence for Example 2.11 using $\alpha = 10I$ and *fsolve* with initial guess $U^{(0)} = 0$.

	$\beta = \mathbf{1}$		$\beta = \mathbf{0}$	
h	Error	Order	Error	Order
2.36e-01	6.80e-01		6.80e-01	
1.77e-01	5.12e-01	0.99	4.60e-01	1.35
1.41e-01	4.01e-01	1.09	3.36e-01	1.41
1.18e-01	3.29e-01	1.09	2.64e-01	1.33
1.01e-01	2.80e-01	1.05	2.18e-01	1.23
8.84e-02	2.44e-01	1.01	1.87e-01	1.14
7.86e-02	2.18e-01	0.98	1.65e-01	1.06

Example 2.12. *Consider the stationary Hamilton-Jacobi-Bellman problem*

$$\begin{aligned} \min\{-\Delta u, -\Delta u/2\} &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega, \end{aligned}$$

where $\Omega = (0, \pi) \times (-\pi/2, \pi/2)$,

$$f(x, y) = \begin{cases} 2\cos(x)\,\sin(y), & \text{if } (x, y) \in S, \\ \cos(x)\,\sin(y), & \text{otherwise,} \end{cases}$$

$S = (0, \pi/2] \times (-\pi/2, 0] \cup (\pi/2, \pi] \times (0, \pi/2)$, *and* g *is chosen such that the viscosity solution is given by* $u(x, y) = \cos(x)\,\sin(y)$.

We can see that the optimal coefficient for Δu varies over four patches in the domain. The approximation results can be found in Table 2.18 and Figure 2.16.

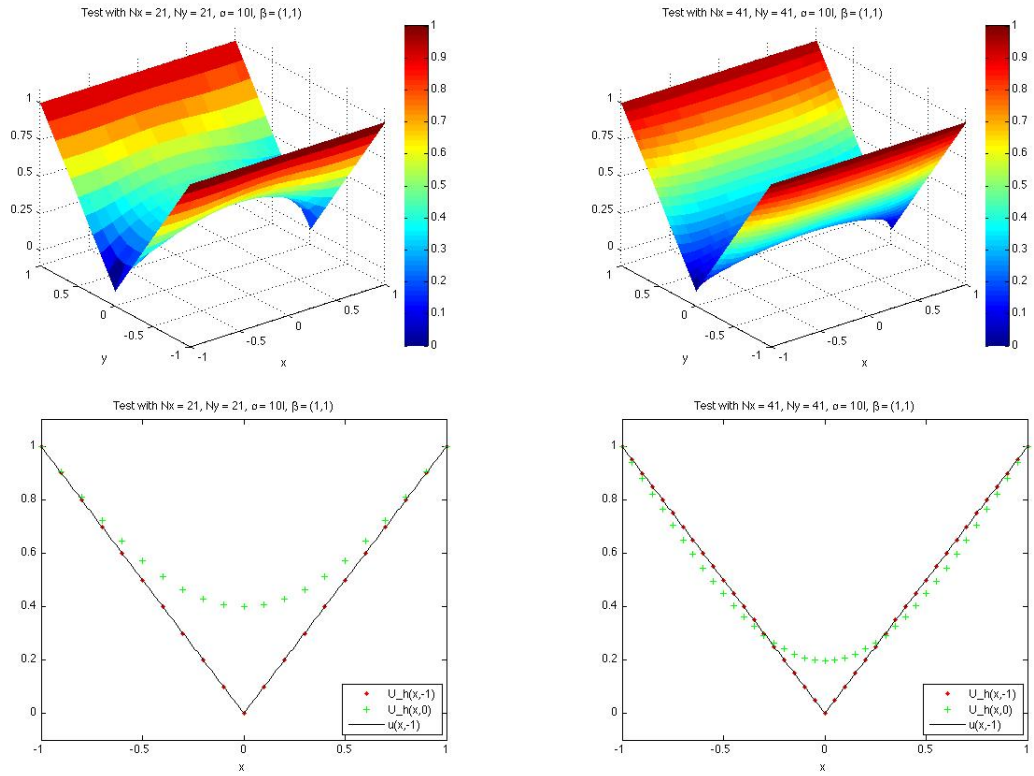


Figure 2.15: Computed solutions for Example 2.11 using $\alpha = 10I$, $\beta = \mathbf{1}$, and *fsolve* with initial guess $U^{(0)} = 0$. The left plots use $h =$ 1.41e-01. The right plots use $h =$ 7.07e-02.

Observe, the rates of convergence appear suboptimal for both schemes, $\beta = 2\mathbf{1}$ and $\beta = \mathbf{0}$. Since u is smooth, the deteriorated convergence rates may be associated with the Hamilton-Jacobi-Bellman operator.

Example 2.13. *Consider the infinite Laplacian problem*

$$\begin{aligned} -\Delta_\infty u := -u_{xx}\,u_x\,u_y - u_{xy}\,u_x\,u_y - u_{yx}\,u_y\,u_y - u_{yy}\,u_y\,u_y = 0 \qquad & \text{in } \Omega, \\ u = g \qquad & \text{on } \partial\Omega, \end{aligned}$$

where $\Omega = (-1,1)\times(-1,1)$ *and* g *is chosen such that the viscosity solution is given by* $u(x,y) = |x|^{4/3} - |y|^{4/3}$. *While this problem is semilinear and not fully nonlinear, the solution has low regularity due to the fact* $u \in C^{1,\frac{1}{3}}(\overline{\Omega}) \cap H^1(\Omega)$.

Table 2.18: Rates of convergence for Example 2.12 using $\alpha = 12I$ and *fsolve* with initial guess $U^{(0)} = 0$.

	$\beta = \mathbf{21}$		$\beta = \mathbf{0}$	
h	Error	Order	Error	Order
1.53e-01	1.83e-01		1.83e-01	
1.35e-01	1.68e-01	0.65	1.62e-01	0.95
1.20e-01	1.56e-01	0.65	1.44e-01	1.03
1.08e-01	1.44e-01	0.76	1.28e-01	1.17
9.87e-02	1.33e-01	0.84	1.14e-01	1.20

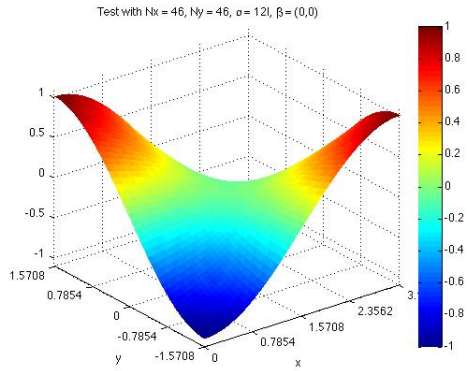


Figure 2.16: Computed solution for Example 2.12 using $\alpha = 12I$, $\beta = \mathbf{0}$, $h =$ 9.87e-02, and *fsolve* with initial guess $U^{(0)} = 0$.

The approximation results can be found in Table 2.19 and Figure 2.17. Observe, the rates of convergence appear to be bounded above by one. Again, the rates appear slightly better for $\beta = \mathbf{0}$. However, the asymptotic rates may be the same due to the lower regularity of the solution.

Table 2.19: Rates of convergence for Example 2.13 using $\alpha = 6I$ and *fsolve* with initial guess $U^{(0)} = 0$.

	$\beta = \mathbf{1}$		$\beta = \mathbf{0}$	
h	Error	Order	Error	Order
1.23e-01	1.02e-01		9.72e-02	
1.05e-01	9.24e-02	0.62	8.68e-02	0.71
9.12e-02	8.43e-02	0.67	7.83e-02	0.74
8.08e-02	7.77e-02	0.67	7.14e-02	0.77
7.25e-02	7.20e-02	0.71	6.55e-02	0.80

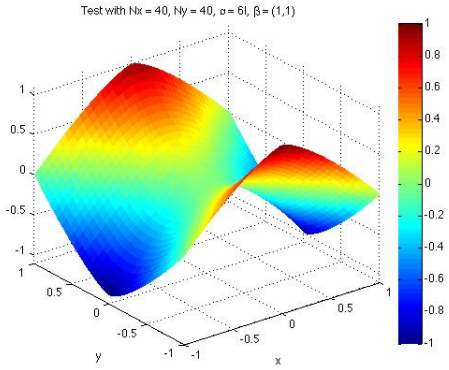


Figure 2.17: Computed solution for Example 2.13 using $\alpha = 6I$, $\beta = \mathbf{1}$, $h = 7.25\text{e-}02$, and *fsolve* with initial guess $U^{(0)} = 0$.

Chapter 3

Local Discontinuous Galerkin Methods

This chapter focuses on the development of a class of discontinuous Galerkin (DG) methods for directly approximating the viscosity solutions of second order Dirichlet boundary value problems

$$F[u](x) := F\left(D^2u, \nabla u, u, x\right) = 0, \qquad x \in \Omega \subset \mathbb{R}^d, \tag{3.1a}$$

$$u(x) = g(x), \qquad x \in \partial\Omega \tag{3.1b}$$

and the viscosity solutions of second order initial boundary value problems

$$u_t + F\left(D^2u, \nabla u, u, x, t\right) = 0, \qquad (x,t) \in \Omega_T := \Omega \times (0,T], \tag{3.2a}$$

$$u(x,t) = g(x), \qquad (x,t) \in \partial\Omega_T := \partial\Omega \times (0,T], \tag{3.2b}$$

$$u(x,0) = u_0(x), \qquad x \in \Omega, \tag{3.2c}$$

where F is a fully nonlinear elliptic operator, Ω is an open, bounded, convex domain, and $T \in \mathbb{R}$ is positive. The methods proposed will generalize the finite difference (FD) methods of Chapter 2 while taking full advantage of the flexibility and versatility of DG methods in terms of accuracy and mesh design. Furthermore, we will show that

the proposed methods posses solver-friendly properties when compared to standard discretization techniques. These solver-friendly properties include more freedom in the choice of the nonlinear solver and less dependence upon the initial guess. The proposed methods will be demonstrated through numerical experiments in Section 3.5.

To introduce a key motivation for our local discontinuous Galerkin (LDG) methods (see [10]), we briefly describe the nonstandard LDG method proposed by Yan and Osher in [53] for approximating the viscosity solution of the Hamilton-Jacobi equation: $u_t + H(\nabla u, u, x, t) = 0$. The main ideas of [53] are to approximate the "left" and "right" side first order derivatives of the viscosity solution and to judiciously combine them through a monotone and consistent numerical Hamiltonian such as the Lax-Friedrichs numerical Hamiltonian (see Section 2.2). We note that the idea of pulling the highest order derivative(s) outside of fully nonlinear PDEs is essential because it allows one to take advantages of DG techniques to discretize the given fully nonlinear PDEs. Our nonstandard LDG methods to be presented below are exactly inspired by this idea, although the realization of this idea for fully nonlinear second order PDEs is more involved. Below we highlight the main steps/ideas of the construction of our nonstandard LDG framework in Section 3.2.

3.1 Notation

To develop our schemes, we first introduce some (standard) notation for DG methods. Let $\mathcal{T}_h$ denote a locally quasi-uniform and shape-regular partition of the domain Ω (see [9]) with $h = \max_{K\in\mathcal{T}_h} \operatorname{diam} K$. Then, we define the following broken H^1-space and broken C^0-space

$$H^1(\mathcal{T}_h) := \prod_{K\in\mathcal{T}_h} H^1(K), \qquad C^0(\mathcal{T}_h) := \prod_{K\in\mathcal{T}_h} C^0(\overline{K}),$$

and the L^2-inner product

$$(v, w)_{\mathcal{T}_h} := \sum_{K \in \mathcal{T}_h} \int_K v\, w\, dx \qquad \forall v, w \in L^2(\mathcal{T}_h).$$

Let $\mathcal{E}_h^I$ denote the set of all interior faces/edges of $\mathcal{T}_h$, $\mathcal{E}_h^B$ denote the set of all boundary faces/edges of $\mathcal{T}_h$, and $\mathcal{E}_h := \mathcal{E}_h^I \cup \mathcal{E}_h^B$. Then, for a set $\mathcal{S}_h \subset \mathcal{E}_h$, we define the broken L^2-inner product over $\mathcal{S}_h$ by

$$\langle v, w \rangle_{\mathcal{S}_h} := \sum_{e \in \mathcal{S}_h} \int_e v\, w\, ds \qquad \forall v, w \in L^2(\mathcal{S}_h).$$

For a fixed integer $r \geq 0$, we define the standard DG finite element space $V^h \subset H^1(\mathcal{T}_h) \subset L^2(\mathcal{T}_h)$ by

$$V^h := \prod_{K \in \mathcal{T}_h} \mathbb{P}_r(K),$$

where $\mathbb{P}_r(K)$ denotes the set of all polynomials on K with degree not exceeding r.

Define the computational domain $\Omega_h := \bigcup_{K \in \mathcal{T}_h} K$. Observe, if Ω is not a polygonal domain, then we have $\overline{\Omega}_h \subsetneq \overline{\Omega}$. Consequently, there exists at least one face/edge $e \in \mathcal{E}_h^B$ such that $e \not\subset \partial\Omega$, and, as a result, special care must be taken to enforce the boundary condition. One solution is to use curvilinear elements, as discussed in [34]. However, we will propose another solution that better suits our methodology and only amounts to a perturbation of the boundary data that should disappear as $h \to 0$. To this end, we introduce a transformation $\sim : \mathcal{E}_h^B \to \partial\Omega$ such that $\sim$ maps each face/edge $e \in \mathcal{E}_h^B$ to a portion of the boundary of the domain, called $\widetilde{e}$, with $\bigcup_{e \in \mathcal{E}_h^B} \widetilde{e} = \partial\Omega$. We also define $\widetilde{\mathcal{E}}_h^B := \{\widetilde{e} \mid e \in \mathcal{E}_h^B\}$.

We can specify the mapping $\sim$ as follows. Let $e \in \mathcal{E}_h^B$, and let X_e denote the set of nodes of e. By the convexity of the domain, we have $X_e \subset \partial\Omega$. First, suppose $d = 2$. Then, we let $\widetilde{e}$ denote the segment of $\partial\Omega$ bounded by the two nodes in X_e. Now, suppose $d \geq 3$. Then, X_e defines a (d-1)-dimensional simplex, and we choose $\widetilde{e} \subset \partial\Omega$ such that

(i) $\bigcup_{e\in\mathcal{E}_h^B} \widetilde{e} = \partial\Omega$.

(ii) $\widetilde{e_1} \cap \widetilde{e_2}$ is a zero-measure set when $e_1 \neq e_2$.

(iii) X_e is the set of all nodes for $\widetilde{e}$.

Note that such a partitioning of $\partial\Omega$ exists.

We now define (standard) interior face/edge dependent functions. Choose $K, K' \in \mathcal{T}_h$, and let $e = \partial K \cap \partial K' \in \mathcal{E}_h^I$. Without loss of generality, we assume that the global labeling number of K is smaller than that of K' and define the following (standard) jump and average notation:

$$[v] := v|_K - v|_{K'}, \qquad \{v\} := \frac{v|_K + v|_{K'}}{2} \tag{3.3}$$

for any $v \in H^m(\mathcal{T}_h)$. We also define $n_e := n_K = -n_{K'}$ as the normal vector to e. For $e \in \mathcal{E}_h^B$, we define n_e as the unit outward normal for the underlying boundary simplex. We note that we will handle function values defined on $\mathcal{E}_h^B$ in a nonstandard way for the DG methods developed in Sections 3.2 and 3.3 that will be based on the boundary mapping $\sim$ and the degree of the polynomial basis r. However, when we only consider the case $r \geq 1$, the boundary function values can be treated in a standard way as in [26].

Lastly, we define the projection operator $\mathcal{P}_h : L^2(\mathcal{T}_h) \to V^h$ by

$$\big(\mathcal{P}_h v, \phi_h\big)_{\mathcal{T}_h} = \big(v, \phi_h\big)_{\mathcal{T}_h} \qquad \forall \phi_h \in V^h \tag{3.4}$$

for all $v \in L^2(\mathcal{T}_h)$. Thus, $\mathcal{P}_h$ is the projection operator onto V^h that is induced by the broken L^2-inner product.

3.2 A Monotone Framework for Second Order Elliptic Problems

In this section, we design a class of DG methods that are based on a nonstandard mixed formulation for problem (3.1) that resembles an LDG methodology. The presented framework modifies the PDE operator F to add strong monotonicity properties as a means for the LDG schemes to select the correct solution to approximate. The DG methodology will give more accurate numerical solutions than the FD framework presented in Chapter 2. We will see in the numerical examples of Section 3.5 that the presented framework appears to successfully remove all numerical artifacts in certain instances.

Several novel ideas are utilized to design direct DG methods for discretizing fully nonlinear PDE problems, all of which will be discussed in the following. The new ideas will provide a way to overcome the inherent difficulties that arise when using Galerkin-based numerical methods to discretize a PDE operator that cannot be expressed in divergence form. We will also see that the DG framework contains schemes that are direct realizations of the vanishing moment methodology found in [29].

3.2.1 Motivation

We now present the five key ideas for our formulation of DG methods for fully nonlinear second order elliptic problems with an underlying viscosity solution in $C^0(\Omega)$. Since integration by parts, which is the necessary tool for constructing any DG method, cannot be performed on equation (3.1a), *the first key idea* is to introduce the auxiliary variables $P := D^2u$ and $q := \nabla u$ and rewrite the original fully nonlinear

PDE as a system of PDEs:

$$F(P, q, u, x) = 0, \tag{3.5a}$$

$$q - \nabla u = 0, \tag{3.5b}$$

$$P - \nabla q = 0. \tag{3.5c}$$

Unfortunately, since ∇u and $D^2 u$ may not exist for a viscosity solution $u \in C^0(\Omega)$, the the above mixed formulation may not make sense. To overcome this difficulty, *the second key idea* is to replace $q := \nabla u$ by two possible values of ∇u, namely, the left and right limits, and $P := \nabla q$ by two possible values for each possible q, namely, the left and right limits. Thus, we have the auxiliary variables $q^+, q^- : \Omega \to \overline{\mathbb{R}}^d$ and $P^{++}, P^{+-}, P^{-+}, P^{--} : \Omega \to \overline{\mathbb{R}}^{d \times d}$ such that

$$q^+(x) - \nabla u(x^+) = 0, \tag{3.6a}$$

$$q^-(x) - \nabla u(x^-) = 0, \tag{3.6b}$$

$$P^{++}(x) - \nabla q^+(x^+) = 0, \tag{3.6c}$$

$$P^{+-}(x) - \nabla q^+(x^-) = 0, \tag{3.6d}$$

$$P^{-+}(x) - \nabla q^-(x^+) = 0, \tag{3.6e}$$

$$P^{--}(x) - \nabla q^-(x^-) = 0. \tag{3.6f}$$

To incorporate the multiple values of ∇u and $D^2 u$, equation (3.5a) must be modified because F is only defined for single value functions P and q. To this end, we need *the third key idea*, which is to replace (3.5a) by

$$\widehat{F}(P^{++}, P^{+-}, P^{-+}, P^{--}, q^+, q^-, u, x) = 0, \tag{3.7}$$

where $\widehat{F}$, which is called a *numerical operator*, should be some well-chosen approximation to F.

The last two key ideas address the criterions for $\widehat{F}$ and how such a numerical operator $\widehat{F}$ can be constructed. Borrowing and adapting the notion of the numerical operators from Chapter 2, where a general FD framework has been developed for fully nonlinear second order PDEs, we arrive at *the fourth key idea*. In short, the criterions for $\widehat{F}$ include *consistency* and *g-monotonicity* (generalized monotonicity), for which precise definitions can be found in Section 2.5. To incorporate the definitions into a DG context, we only need to extend the definition of g-monotonicity to all $x \in \Omega$.

Finally, we need to design a DG discretization for the mixed system (3.6) and (3.7) that incorporates all of the above key ideas. To this end, *the fifth key idea* is to use different *numerical fluxes* in the formulations of DG methods for the "one-sided" linear problems represented by (3.6) in concert with the addition of a *numerical moment*, which can be regarded as a direct numerical realization for the moment term in *the vanishing moment methodology* introduced in [29].

The consistency and g-monotonicity of the numerical operator play a critical role in the LDG methods we formulate for fully nonlinear second order PDE problems. Using a numerical moment, we can immediately design numerical operators that fulfill the consistency and g-monotonicity requirements. To this end, we propose the Lax-Friedrichs-like numerical operator first introduced in Chapter 2:

$$\begin{aligned} &\widehat{F}\left(P^{++}, P^{+-}, P^{-+}, P^{--}, q^{+}, q^{-}, v, x\right) \\ &\qquad := F\left(\frac{P^{+-}+P^{-+}}{2}, \frac{q^{+}+q^{-}}{2}, v, x\right) \\ &\qquad\quad + \alpha : \left(P^{++} - P^{+-} - P^{-+} + P^{--}\right) - \beta \cdot \left(q^{+} - q^{-}\right), \end{aligned} \tag{3.8}$$

where $\alpha \in \mathbb{R}^{d\times d}$ is a nonnegative constant matrix and $\beta \in \mathbb{R}^{d}$ is a nonnegative constant vector chosen to enforce the g-monotonicity property. The second to last term in (3.8) is called *the numerical moment* and the last term in (3.8) is called *the*

numerical viscosity. Notationally, we have

$$A : B := \sum_{i=1}^{d} \sum_{j=1}^{d} A_{i,j} \, B_{i,j}$$

for $A, B \in \mathbb{R}^{d \times d}$.

Suppose F is differentiable and there exists positive constants M_α and M_β such that

$$M_\alpha > \left\| \frac{\partial F}{\partial D^2 u} \right\|_{L^\infty(\mathbb{R}^{d \times d})}, \qquad M_\beta > \left\| \frac{\partial F}{\partial \nabla u} \right\|_{L^\infty(\mathbb{R}^d)}. \tag{3.9}$$

Then, a sufficient choice for α and β that yields g-monotonicity is given by

$$\alpha = d \, M_\alpha \, I, \qquad \beta = d \, M_\beta, \vec{1}, \tag{3.10}$$

where I denotes the identity matrix for $\mathbb{R}^{d \times d}$ and $\vec{1}$ denotes the vector with all components equal to one. It is trivial to check using Gershgorin's circle theorem that for the given value of α, $\frac{\partial \widehat{F}}{\partial P^{--}}$ and $\frac{\partial \widehat{F}}{\partial P^{++}}$ are positive definite and $\frac{\partial \widehat{F}}{\partial P^{-+}}$ and $\frac{\partial \widehat{F}}{\partial P^{+-}}$ are negative definite. Furthermore, $\widehat{F}$ is increasing with respect to each component of q^- and decreasing with respect to each component of q^+. Thus, the Lax-Friedrichs-like numerical operator is g-monotone using the choices $\alpha = d \, M_\alpha \, I$ and $\beta = d \, M_\beta \, \vec{1}$.

Remark 3.1.

(a) Due to the definition of ellipticity for F, the g-monotonicity constraints on $\widehat{F}$ with respect to P^{-+} and P^{+-} are natural.

(b) By choosing the numerical moment correctly, the numerical operator $\widehat{F}$ will behave like a uniformly elliptic operator, even if the PDE operator F is a degenerate elliptic operator. The consistency assumption then guarantees that the numerical operator is still a reasonable approximation for the PDE operator.

(c) *While it may not be possible to globally bound* $\frac{\partial F}{\partial D^2 u}$*, it may be sufficient to choose a value for* α *such that the g-monotonicity property is preserved locally over each iteration of the nonlinear solver for a given initial guess.*

(d) *It may be beneficial for a given nonlinear solver to let* $\alpha = d\,M\,I + M\,\mathbf{1}$*, where* $\mathbf{1} \in \mathbb{R}^{d\times d}$ *such that* $\mathbf{1}_{i,j} = 1$ *for all* $i, j = 1, 2, \ldots, d$*, to ensure that* $\widehat{F}$ *is also monotone with respect to each component of* $P^{\mu\nu}$ *for* $\mu, \nu \in \{-, +\}$*.*

(e) *The numerical moment can be used to design a modified fixed-point solver for the resulting nonlinear system of equations that will be formulated in the sections that follow. The solver will be defined in Section 3.4.1.*

(f) *The role of the numerical moment will be discussed further in Section 3.5.3.*

3.2.2 Element-Wise Formulation of the LDG Methods

We now develop an element-wise formulation for our LDG methods. First we introduce some local definitions. Let $K \in \mathcal{T}_h$ and n denote the unit outward normal vector to ∂K. We also let $K' \in \mathcal{T}_h$ such that the set $e = \partial K \cap \partial K'$ forms a face/edge in the triangulation. Then, for all functions $v \in V^h$, let $v\left(x^I\right)$ denote the value of $v(x)$ on ∂K from the interior of the element K and $v\left(x^E\right)$ denote the value of $v(x)$ on ∂K from the interior of the element K'. Using these limit definitions, we define the local boundary flux operators $T^+, T^- : \mathcal{P}_r(K) \to \left(\prod_{e\subset\partial K} \mathcal{P}_r(e)\right)^d$ by

$$T_i^-(v)(x) := \begin{cases} v\left(x^I\right), & \text{if } n_i(x) \geq 0, \\ v\left(x^E\right), & \text{otherwise,} \end{cases} \tag{3.11a}$$

$$T_i^+(v)(x) := \begin{cases} v\left(x^E\right), & \text{if } n_i(x) \geq 0, \\ v\left(x^I\right), & \text{otherwise} \end{cases} \tag{3.11b}$$

for all $i \in \{1, 2, \ldots, d\}$, $x \in e$, and $v \in V^h$. We will define values for $T_i^{\pm}(v)(x)$ when $v \in V^h$ and $x \in e$ for some $e \in \mathcal{E}_h^B$ in Section 3.2.4.

We are now ready to formulate our DG discretizations for the system of equations given by (3.6) and (3.7). First, we approximate the (fully) nonlinear equation (3.7) by simply using its broken L^2-projection into V^h, namely,

$$a_0\big(u_h, q_h^+, q_h^-, P_h^{++}, P_h^{+-}, P_h^{-+}, P_h^{--}; \phi_{0h}\big) = 0 \qquad \forall \phi_{0h} \in V^h, \tag{3.12}$$

where

$$\begin{aligned} &a_0(u, q^+, q^-, P^{++}, P^{+-}, P^{-+}, P^{--}; \phi_0) \\ &\quad = \Big(\widehat{F}(P^{++}, P^{+-}, P^{-+}, P^{--}, q^+, q^-, u, \cdot), \phi_0\Big)_{\mathcal{T}_h}. \end{aligned}$$

Next, we discretize the six *linear* equations (3.6) locally with respect to each component using the partial derivative identity

$$\int_S v_{x_i}\, \varphi \, dx = \int_{\partial S} v\, \varphi\, n_i \, ds - \int_S v\, \varphi_{x_i}\, dx \qquad \forall \varphi \in C^1(S) \tag{3.13}$$

for $i = 1, 2, \ldots, d$. Thus, the above identity yields an integral representation for the partial derivative v_{x_i} on the set S for all $v \in H^1(S)$. Using the preceding identity, we define our gradient approximations $q_h^\mu \in \big(V^h\big)^d$, $\mu \in \{+, -\}$, by

$$\int_K q_i^\mu\, \phi_i^\mu \, dx + \int_K u\, \left(\phi_i^\mu\right)_{x_i}\, dx = \int_{\partial K} T_i^\mu(u)\, n_i\, \phi_i^\mu\left(x^I\right)\, ds \qquad \forall \phi_i^\mu \in V^h \tag{3.14}$$

for $i = 1, 2, \ldots, d$, $\mu = +, -$. Similarly, we define our Hessian approximations $P_h^{\mu\,\nu} \in \big(V^h\big)^{d\times d}$, $\mu, \nu \in \{+, -\}$, by

$$\int_K P_{i,j}^{\mu\,\nu}\, \psi_{i,j}^{\mu\,\nu}\, dx + \int_K q_i^\mu\, \left(\psi_{i,j}^{\mu\,\nu}\right)_{x_j}\, dx = \int_{\partial K} T_j^\nu\left(q_i^\mu\right)\, n_j\, \psi_{i,j}^{\mu\,\nu}\left(x^I\right)\, ds \qquad \forall \psi_{i,j}^{\mu\,\nu} \in V^h \tag{3.15}$$

for $i, j = 1, 2, \ldots, d$, $\mu, \nu = +, -$.

Thus, in order to approximate the viscosity solution u for the fully nonlinear Dirichlet boundary value problem (3.1a), we seek functions $u_h \in V^h$; q_h^+,$q_h^- \in [V^h]^d$;

and P_h^{++}, P_h^{+-}, P_h^{-+}, $P_h^{--} \in [V^h]^{d\times d}$ such that equation (3.12) holds as well as equations (3.14) and (3.15) for all $K \in \mathcal{T}_h$, where u_h forms the approximation for u.

3.2.3 Whole Domain Formulation of the LDG Methods

From above, we see that the local boundary flux operators are essential for developing an element-wise DG formulation. Choose $e \in \mathcal{E}_h$. Then, there exists an element $K \in \mathcal{T}_h$ such that $n_K = n_e$. Thus, we can extend the local boundary flux operators $T^{\pm}$ from Section 3.2.2 to $\mathcal{E}_h$ by evaluating $T^{\pm}(v)(x)$ for $x \in e$ using the local definition (3.11) for K.

Summing equations (3.14) and (3.15) from the element-wise formulation over all elements $K \in \mathcal{T}_h$, we have the following whole-domain DG discretization of (3.6):

$$\left(q_i^{\mu}, \phi_i^{\mu}\right)_{\mathcal{T}_h} + a_i^{\mu}\left(u_h, \phi_i^{\mu}\right) = 0 \qquad \forall \phi_i^{\mu} \in V^h, \tag{3.16a}$$

$$\left(P_{i,j}^{\mu\,\nu}, \psi_{i,j}^{\mu\,\nu}\right)_{\mathcal{T}_h} + a_j^{\nu}\left(q_i^{\mu}, \psi_{i,j}^{\mu\,\nu}\right) = 0 \qquad \forall \psi_{i,j}^{\mu\nu} \in V^h \tag{3.16b}$$

for $i, j = 1, 2, \ldots, d$, $\mu, \nu = -, +$, where

$$a_i^{\mu}(v, \phi) := \left(v, \phi_{x_i}\right)_{\mathcal{T}_h} - \left\langle T_i^{\mu}(v), [\phi]\; n_{e_i}\right\rangle_{\mathcal{E}_h^I} - \left\langle T_i^{\mu}(v), \phi(x^I)\, n_i\right\rangle_{\mathcal{E}_h^B}$$

for all $v, \phi \in V^h$ and $i, j = 1, 2, \ldots, d$, $\mu, \nu = +, -$. We note that the boundary condition will appear as a constraint equation based upon the choice of boundary flux values that will be defined in the following section.

In summary, our nonstandard LDG methods for the fully nonlinear Dirichlet boundary value problem (3.1a) are defined as seeking $\left(u_h, q_h^+, q_h^-, P_h^{++}, P_h^{+-}, P_h^{-+}, P_h^{--}\right) \in V^h \times \left([V^h]^d\right)^2 \times \left([V^h]^{d\times d}\right)^4$ such that (3.12) and (3.16) hold.

3.2.4 Boundary Flux Values

We now extend the definition for the boundary flux operators, given by (3.11), to the set $\mathcal{E}_h^B$. To this end, we will introduce a set of constraint equations that expresses all

exterior limits in terms of interior limits and known data. The Dirichlet boundary data will serve as an exterior constraint on the sought-after numerical solution. We will consider two cases based on the definition of V^h: $r \geq 1$ and $r = 0$. For $r \geq 1$, we will enforce a "continuity" assumption across the boundary. For $r = 0$, we will prescribe an alternative approach that will more closely resemble the introduction of "ghost values" commonly used in FD methods.

Prior to introducing the constraint equations, we specify a convention to be used for all boundary faces/edges. Let $K \in \mathcal{T}_h$ be a boundary simplex, and let $e \in \mathcal{E}_h^B$ such that $e \subset \partial K$. Suppose $v_h \in V^h$ such that v_h is supported on K. Then, we define $v_h(x) := v_h(x^I)$ for all $x \in e$. Furthermore, we may have $\overline{\Omega}_h \subsetneq \overline{\Omega}$. In this case, we simply extend the support of v_h to include $\widetilde{e}$ (see Section 3.1). Such an extension is trivial since v_h is a polynomial.

We first consider $r \geq 1$, in which case we make the "continuity" assumption

$$v_h(x^E) = v_h(x) \tag{3.17}$$

for all $x \in e$ and $v_h \in V^h$ such that $e \in \mathcal{E}_h^B$. Since problem (3.1) lacks a Neumman boundary condition, we simply treat $q_j^{\pm}(x)$ as an unknown for all $j = 1, 2, \ldots, d$ and $x \in e$ such that $e \in \mathcal{E}_h^B$. Alternatively, when defining the boundary flux values for u_h, we use the Dirichlet boundary condition given by (3.1b). Thus, for $r \geq 1$, we wish to impose

$$u_h(x) = g(x)$$

for all $x \in \partial\Omega$. However, g may not be a polynomial of degree r, and we may have $\overline{\Omega}_h \subsetneq \overline{\Omega}$. Thus, we introduce the constraint equations

$$\sum_{i=1}^{d} \left\langle u_h(x), \varphi_h(x) n_i \right\rangle_{\mathcal{E}_h^B} = \sum_{i=1}^{d} \left\langle g(x), \varphi_h(x) n_i \right\rangle_{\widetilde{\mathcal{E}}_h^B} \qquad \forall \varphi_h \in V^h \tag{3.18}$$

using the boundary extension operator defined in Section 3.1, where n denotes the unit outward normal vector. The constraint equation will fix some of the degrees of

freedom in our approximation space so that our approximation satisfies the boundary condition in a weak sense. Observe, when a boundary simplex has more than one face/edge in $\mathcal{E}_h^B$, we are treating all of the boundary simplex's faces/edges in $\mathcal{E}_h^B$ as a single (d-1)-dimensional surface.

We now consider the case $r = 0$ for the remainder of the section. Extending the definition for the boundary flux operators, given by (3.11), to the set $\mathcal{E}_h^B$ is less straightforward for the special case $r = 0$. We can see this by observing the fact that when fixing the interior limit of a boundary value on a boundary simplex, we actually fix the function value on the entire simplex. Thus, strictly enforcing a Dirichlet boundary condition for u_h may result in a boundary layer with respect to the overall approximation error when measured in low-order norms such as the L^∞ norm or L^2 norm. Our goal is to prescribe boundary flux values in a way that results in a potential boundary layer that corresponds to only high-order error, i.e., boundary layers that only appear when measuring the approximation error in $W^{1,\infty}$ or H^1 semi-norms, when defined.

In order to motivate our choice of boundary flux values when using $r = 0$, we first show how the above LDG formulation relates to the FD methods formulated in Chapter 2. Assume $\Omega \subset \mathbb{R}^d$ is a d-rectangle, i.e., $\Omega = (a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_d, b_d)$. Let $N_1, N_2, \ldots, N_d$ be positive integers, and let $h_i = (b_i - a_i)/N_i$ for all $i = 1, 2, \ldots, d$. Then, we partition the x_i coordinate direction by dividing the interval (a_i, b_i) into N_i subintervals of length h_i for all $i = 1, 2, \ldots, d$, and we let X_{h_i} denote the collection of intervals. Thus, we can define the partition of Ω by $\mathcal{T}_h = \prod_{i=1}^d X_{h_i}$, where $h = (h_1, h_2, \ldots, h_d)$ and $|\mathcal{T}_h| = \prod_{i=1}^d N_i$. Lastly, we label the elements of $\mathcal{T}_h$ using the natural ordering.

Letting $r = 0$ in the above LDG formulation implies that the approximation u_h is constant on each rectangle in $\mathcal{T}_h$. Let x_K denote the center of K for all $K \in \mathcal{T}_h$, and let ξ_i denote the unit vector in the Cartesian direction of x_i for all $i = 1, 2, \ldots, d$. Then, $[x_K]_i = a_i + (\alpha_i - 1/2)\, h_i$ for some $\alpha_i \in \{1, 2, \ldots, N_i\}$. Thus, each node x_K can be fully described by a multi-index α_K.

Define the grid function U by $U_{\alpha_K} = u_h(x_{\alpha_K})$ for all $K \in \mathcal{T}_h$. Pick $i, j \in \{1, 2, \ldots, d\}$ and let $K \in \mathcal{T}_h$ be an interior simplex. Then, letting $\phi_i^{\pm} = \chi_K$ in (3.14) for χ_K the characteristic function on K, we have

$$q_i^-(x_{\alpha_K}) = \frac{1}{h_i}\left(U_{\alpha_K} - U_{\alpha_K - \xi_i}\right) = \delta^-_{x_i,h_i} U_{\alpha_K}, \tag{3.19a}$$

$$q_i^+(x_{\alpha_K}) = \frac{1}{h_i}\left(U_{\alpha_K + \xi_i} - U_{\alpha_K}\right) = \delta^+_{x_i,h_i} U_{\alpha_K}, \tag{3.19b}$$

where $\delta^{\pm}_{x_i,h_i}$ denotes the standard forward and backward difference operator in FD. Similarly, letting $\psi_{i,j}^{\mu\nu} = \chi_K$ in (3.15) for $\mu, \nu \in \{-, +\}$, we have

$$\begin{aligned} P_{i,j}^{--}(x_{\alpha_K}) &= \frac{1}{h_j}\left(q_i^-(x_{\alpha_K}) - q_i^-(x_{\alpha_K} - h_j\xi_j)\right) && (3.20a)\\ &= \frac{1}{h_i h_j}\left(U_{\alpha_K} - U_{\alpha_K-\xi_i} - U_{\alpha_K-\xi_j} + U_{\alpha_K-\xi_j-\xi_i}\right) = \delta^-_{x_i,h_i}\delta^-_{x_j,h_j} U_{\alpha_K}, \\ P_{i,j}^{-+}(x_{\alpha_K}) &= \frac{1}{h_j}\left(q_i^-(x_{\alpha_K} + h_j\xi_j) - q_i^-(x_{\alpha_K})\right) && (3.20b)\\ &= \frac{1}{h_i h_j}\left(U_{\alpha_K+\xi_j} - U_{\alpha_K+\xi_j-\xi_i} - U_{\alpha_K} + U_{\alpha_K-\xi_i}\right) = \delta^-_{x_i,h_i}\delta^+_{x_j,h_j} U_{\alpha_K}, \\ P_{i,j}^{+-}(x_{\alpha_K}) &= \frac{1}{h_j}\left(q_i^+(x_{\alpha_K}) - q_i^+(x_{\alpha_K} - h_j\xi_j)\right) && (3.20c)\\ &= \frac{1}{h_i h_j}\left(U_{\alpha_K+\xi_i} - U_{\alpha_K} - U_{\alpha_K-\xi_j+\xi_i} + U_{\alpha_K-\xi_j}\right) = \delta^+_{x_i,h_i}\delta^-_{x_j,h_j} U_{\alpha_K}, \\ P_{i,j}^{++}(x_{\alpha_K}) &= \frac{1}{h_j}\left(q_i^+(x_{\alpha_K} + h_j\xi_j) - q_i^+(x_{\alpha_K})\right) && (3.20d)\\ &= \frac{1}{h_i h_j}\left(U_{\alpha_K+\xi_j+\xi_i} - U_{\alpha_K+\xi_j} - U_{\alpha_K+\xi_i} + U_{\alpha_K}\right) = \delta^+_{x_i,h_i}\delta^+_{x_j,h_j} U_{\alpha_K}. \end{aligned}$$

Hence, for the case $r = 0$ on a uniform rectangular grid, we successfully recover the FD methods formulated in Chapter 2 for the interior of the domain, and it follows that by extending the equivalence of the two methods to the boundary of the domain, we can derive the necessary boundary flux values for u_h and $q_h^{\pm}$ on $\mathcal{E}_h^B$.

In order to define the boundary values for u_h, q_h^+, and q_h^-, we will need to develop a methodology for extending the solution u to the exterior of the domain Ω. We now define a way to do such an extension that is consistent with the interpretation of the

auxiliary variables and consistent with the FD strategy of introducing "ghost values" for the grid function U.

We first describe the extension for the approximation function u_h. Given the Dirichlet boundary data for the viscosity solution u, it is natural to assume that the approximation function u_h has a constant extension beyond each individual boundary face/edge. Thus, we wish to define the exterior boundary fluxes using the Dirichlet boundary condition (3.1b) by setting

$$u\left(x^E\right) = g(x)$$

for all $x \in \partial K \cap \partial\Omega$. However, Ω may not be polygonal, and a given boundary simplex may have multiple faces/edges in $\mathcal{E}_h^B$. Therefore, we introduce a "ghost simplex" exterior to each individual face/edge in $\mathcal{E}_h^B$, and define the exterior value as g_e, where

$$\sum_{i=1}^{d} \langle g_e, n_i \rangle_e = \sum_{i=1}^{d} \langle g, n_i \rangle_{\widetilde{e}} \tag{3.21}$$

for each $e \in \mathcal{E}_h^B$. Then, we define

$$u_h(x^E)\Big|_e := g_e \tag{3.22}$$

for all faces/edges $e \in \mathcal{E}_h^B$.

Observe that for $r = 0$ we only apply the Dirichlet boundary condition to the exterior function limits. Furthermore, we define the exterior function limits to be edge dependent. Since the function value is constant on each simplex K, we do not extend the Dirichlet boundary condition to the interior of the domain by enforcing (3.18). Instead, we treat the value of u_h on K as an unknown whenever K is a boundary simplex. We use the edge dependent definition to mimic the use of ghost values for $r = 0$, which are introduced for each coordinate direction when using a FD methodology. When $\mathcal{T}_h$ is a Cartesian partition, our methodology does in fact result in the introduction of a fixed exterior boundary flux value for each individual coordinate

direction. The result of the methodology will be a more weighted approximation on a boundary simplex based upon the boundary condition along each boundary face/edge independently and on the PDE for the interior of the simplex.

We now describe how we assign boundary values for $q_h^{\pm}$. Since we do not have Neumman boundary data, we will have to enforce auxiliary boundary conditions. From equations (3.19), we see that

$$q_i^-(x_{\alpha_K}) = q_i^+(x_{\alpha_K - \xi_i}), \qquad q_i^+(x_{\alpha_K}) = q_i^-(x_{\alpha_K + \xi_i}) \tag{3.23}$$

for all $i = 1, 2, \ldots, d$ and all interior simplexes $K \in \mathcal{T}_h$. Let n_e denote the unit normal vector for each $e \in \mathcal{E}_h^B$. Extending (3.23) to the boundary yields

$$q_i^-(x^E) = q_i^+(x^I), \qquad \text{if } n_{e_i} < 0, \tag{3.24a}$$

$$q_i^+(x^E) = q_i^-(x^I), \qquad \text{if } n_{e_i} \geq 0 \tag{3.24b}$$

for $x \in e$, where both $q_i^+(x^I)$ and $q_i^-(x^I)$ are treated as unknowns.

Observe that the above extension does not define exterior limits for q_i^+ if $n_{e_i} < 0$ or $q_{e_i}^-$ if $n_{e_i} \geq 0$. In order to define the remaining exterior limit values, we also impose the auxiliary constraint equations

$$\sum_{i=1}^{d} \left\langle q_i^- \left(x^I\right) - q_i^- \left(x^E\right), n_{e_i} \right\rangle_e = 0, \tag{3.25a}$$

$$\sum_{i=1}^{d} \left\langle q_i^+ \left(x^I\right) - q_i^+ \left(x^E\right), n_{e_i} \right\rangle_e = 0 \tag{3.25b}$$

for each face/edge $e \in \mathcal{E}_h^B$.

The above constraint equations are consistent with discretizing the higher order auxiliary constraint for all ghost-values of $q_h^{\pm}$:

$$\sum_{k=1}^{d} \left(q_k^{\pm}\right)_{x_k} (\widehat{x}) = 0,$$

where $\widehat{x} \in \Omega^c$. The philosophy for such an auxiliary assumption can be found in [29]. We note that the constraint equations (3.25) are also trivially satisfied when defining the exterior values for $r \geq 1$ due to our "continuity" assumption. Assuming that $\mathcal{T}_h$ is either a uniform Cartesian partition or a d-triangular partition where each simplex has at most one face/edge in $\mathcal{E}_h^B$, we can see that all exterior limits on the boundary of the domain have now been expressed in terms of unknown interior limits that correspond to degrees of freedom for the discretization.

Remark 3.2. *When $r = 0$, our approximation space consists of totally discontinuous piecewise constant functions. We have prescribed a way to assign all exterior boundary flux values for our approximation functions, and, by convention, we treat all interior boundary flux values as unknowns.*

3.2.5 Analysis of the Auxiliary Linear Equations

We use this section to further explore the auxiliary system of linear equations. Based upon the boundary flux values defined in Section 3.2.4, we can immediately form linear operators that map u_h to each auxiliary variable q_h^μ, $P_h^{\mu\nu}$ for $\mu, \nu \in \{+, -\}$ (see Section 3.3). A complete analysis of these linear operators presented as discrete derivative operators can be found in [26]. In this section, we instead focus on the inverse problem that involves mapping a "discrete Laplacian" given by the trace of $(P_h^{+-} + P_h^{-+})/2$ back to the unknown function u_h. In the following, we let $\Lambda_h \in V^h$ be defined by

$$\Lambda_h := \operatorname{tr}\left(\frac{P_h^{+-} + P_h^{-+}}{2}\right), \tag{3.26}$$

where tr denotes the matrix trace operator.

Lemma 3.1. *Suppose $r \geq 1$ in the definition of V^h. Furthermore, assume at least one boundary simplex in $\mathcal{T}_h$ has only one face/edge in $\mathcal{E}_h^B$. Then, using (3.16), we can uniquely determine values for u_h, q_h^+, and q_h^- when given a value for Λ_h, where Λ_h is defined by (3.26).*

Proof. Observe, (3.16) is a linear system of equations. Since V^h is finite-dimensional, it is sufficient to show that if $\Lambda_h = 0$ and $g = 0$ in (3.1b), then $u_h = 0$ and $q_h^+ = q_h^- = 0$.

Pick $i \in \{1, 2, \ldots, d\}$. Letting $\phi_i^\pm = q_i^\pm$ in (3.16a) and using the boundary flux values defined in Section 3.2.4, we have

$$\begin{aligned}
0 &= \left(q_i^\pm, q_i^\pm\right)_{\mathcal{T}_h} + \left(u_h, (q_i^\pm)_{x_i}\right)_{\mathcal{T}_h} - \left\langle T_i^\pm(u_h), \left[q_i^\pm\right] n_{e_i}\right\rangle_{\mathcal{E}_h^I} - \left\langle T_i^\pm(u_h), q_i^\pm(x^I)\, n_i\right\rangle_{\mathcal{E}_h^B} \\
&= \left(q_i^\pm, q_i^\pm\right)_{\mathcal{T}_h} - \left((u_h)_{x_i}, q_i^\pm\right)_{\mathcal{T}_h} + \left\langle u_h(x^I), q_i^\pm(x^I) n_i\right\rangle_{\mathcal{E}_h^B} + \left\langle \left[u_h q_i^\pm\right], n_{e_i}\right\rangle_{\mathcal{E}_h^I} \\
&\quad - \left\langle T_i^\pm(u_h), \left[q_i^\pm\right] n_{e_i}\right\rangle_{\mathcal{E}_h^I} - \left\langle T_i^\pm(u_h), q_i^\pm(x^I)\, n_i\right\rangle_{\mathcal{E}_h^B} \\
&= \left(q_i^\pm, q_i^\pm\right)_{\mathcal{T}_h} - \left((u_h)_{x_i}, q_i^\pm\right)_{\mathcal{T}_h} + \left\langle T_i^\mp(q_i^\pm), [u_h]\, n_{e_i}\right\rangle_{\mathcal{E}_h^I}.
\end{aligned}$$

Also, letting $j = i$, $\mu \neq \nu$, and $\psi_{i,j}^{\mu\nu} = u_h$ in (3.16b), we have

$$0 = \left(P_{i,i}^{\pm\mp}, u_h\right)_{\mathcal{T}_h} + \left(q_i^\pm, (u_h)_{x_i}\right)_{\mathcal{T}_h} - \left\langle T_i^\mp(q_i^\pm), [u_h]\, n_{e_i}\right\rangle_{\mathcal{E}_h^I} - \left\langle T_i^\mp(q_i^\pm), u_h(x^I)\, n_i\right\rangle_{\mathcal{E}_h^B}.$$

Then, summing the above two equations, we have

$$0 = \left(q_i^\pm, q_i^\pm\right)_{\mathcal{T}_h} + \left(P_{i,i}^{\pm\mp}, u_h\right)_{\mathcal{T}_h} - \left\langle T_i^\mp(q_i^\pm), u_h(x^I)\, n_i\right\rangle_{\mathcal{E}_h^B}.$$

Hence, summing over i, we have

$$0 = \left(q_h^+, q_h^+\right)_{\mathcal{T}_h} + \left(q_h^-, q_h^-\right)_{\mathcal{T}_h} + 2\left(\Lambda_h, u_h\right)_{\mathcal{T}_h} - \sum_{i=1}^{d} \left\langle T_i^-(q_i^+) + T_i^+(q_i^-), u_h(x^I)\, n_i\right\rangle_{\mathcal{E}_h^B}.$$

Therefore, using the assumption that $\Lambda_h = 0$, the boundary flux values defined in Section 3.2.4, and the assumptions on $\mathcal{T}_h$, we have $q_h^\pm = 0$.

We now show that u_h is continuous. Observe, letting $\phi_i^\pm = u_h$ in (3.16a), we have

$$(u_h, (u_h)_{x_i})_{\mathcal{T}_h} - \left\langle T_i^\pm(u_h), [u_h]\ n_{e_i}\right\rangle_{\mathcal{E}_h^I} - \left\langle T_i^\pm(u_h), u_h(x^I)\, n_i\right\rangle_{\mathcal{E}_h^B} = 0.$$

Then, we have

$$\begin{aligned}
0 &= \sum_{i=1}^{d} \left\langle T_i^+(u_h) - T_i^-(u_h), [u_h]\ n_{e_i} \right\rangle_{\mathcal{E}_h^I} + \sum_{i=1}^{d} \left\langle T_i^+(u_h) - T_i^-(u_h), u_h(x^I)\, n_i \right\rangle_{\mathcal{E}_h^B} \\
&= \sum_{i=1}^{d} \left\langle [u_h] \operatorname{sgn}(n_{e_i}), [u_h]\ n_{e_i} \right\rangle_{\mathcal{E}_h^I} \\
&= \sum_{i=1}^{d} \left\langle [u_h], [u_h]\ |n_{e_i}| \right\rangle_{\mathcal{E}_h^I},
\end{aligned}$$

where

$$\operatorname{sgn}(a) = \begin{cases} 1 & \text{if } a \geq 0, \\ -1 & \text{if } a < 0. \end{cases}$$

Thus, $\left[u_h\right] = 0$ on $\mathcal{E}_h^I$, and, since u_h is a piecewise polynomial, it follows that $u_h \in H^1(\Omega) \cap C(\Omega)$.

We now show that u_h is constant valued on each simplex. Using (3.16a), we have

$$\begin{aligned}
0 &= (u_h, (\phi_h)_{x_i})_{\mathcal{T}_h} - \left\langle T_i^{\pm}(u_h), [\phi_h]\ n_{e_i} \right\rangle_{\mathcal{E}_h^I} - \left\langle T_i^{\pm}(u_h), \phi_h(x^I)\, n_i \right\rangle_{\mathcal{E}_h^B} \\
&= -\left((u_h)_{x_i}, \phi_h\right)_{\mathcal{T}_h} + \left\langle T_i^{\mp}(\phi_h), [u_h]\ n_{e_i} \right\rangle_{\mathcal{E}_h^I} - \left\langle T_i^{\pm}(u_h), \phi_h(x^I)\, n_i \right\rangle_{\mathcal{E}_h^B} \\
&= -\left((u_h)_{x_i}, \phi_h\right)_{\mathcal{T}_h} - \left\langle T_i^{\pm}(u_h), \phi_h(x^I)\, n_i \right\rangle_{\mathcal{E}_h^B}.
\end{aligned}$$

Choosing $\phi_h = (u_h)_{x_i} \in V^h$, we have

$$\left((u_h)_{x_i}, (u_h)_{x_i}\right)_{\mathcal{T}_h} + \left\langle T_i^{\pm}(u_h), n_i\, (u_h)_{x_i} \right\rangle_{\mathcal{E}_h^B} = 0.$$

Thus, summing over $i = 1, 2, \ldots, d$, we have $\nabla u_h = 0$ on Ω, and it follows that u_h is constant valued on each simplex.

Finally, we show $u_h = 0$. From above, we have u_h is constant valued on each simplex and continuous. Thus, u_h is a constant over Ω. Let $K \in \mathcal{T}_h$ be a boundary simplex with only one face/edge in $\mathcal{E}_h^B$. Let $e = \partial K \cap \mathcal{E}_h^B$ and $\phi_h = \chi_{\overline{K}}\, u_h\, n_i \in V^h$ in

(3.16a), where χ_A denotes the indicator function over the set A. Then,

$$\begin{aligned} 0 &= -\big((u_h)_{x_i}, \phi_h\big)_{\mathcal{T}_h} - \big\langle u_h(x^I), \phi_h(x^I)\, n_i \big\rangle_{\mathcal{E}_h^B} \\ &= -\big\langle u_h(x^I), \phi_h(x^I)\, n_i \big\rangle_{\mathcal{E}_h^B} \\ &= -\big\langle u_h(x^I), u_h(x^I)\, (n_i)^2 \big\rangle_e \,. \end{aligned}$$

Summing over $i = 1, 2, \ldots, d$, we have $u_h = 0$ on K. Therefore, $u_h = 0$. The proof is complete. □

We also have the following result that follows directly from the derivation of the local boundary flux values defined in Section 3.2.4 and the fact that the central difference approximation for the Laplace operator, Λ_h^2, is invertible when given boundary data, i.e., values for $u_h(x^E)$ along each face/edge in $\mathcal{E}_h^B$:

Lemma 3.2. *Suppose $r = 0$ in the definition of V^h. Assume $\mathcal{T}_h$ is a uniform Cartesian mesh. Then, using (3.16), we can uniquely determine values for u_h, q_h^+, and q_h^- when given a value for Λ_h.*

3.2.6 The Numerical Viscosity and the Numerical Moment

In this section, we take a closer look at the numerical viscosity and the numerical moment used in the definition of the Lax-Friedrichs-like numerical operator, (3.8). We will divide the analysis into two cases, $r = 0$ and $r \geq 1$. When $r = 0$, we will see that we recover the numerical viscosity and the numerical moment introduced in Chapter 2 for FD methods. When $r \geq 1$, we will recover interior jump/stabilization terms.

Suppose $r = 0$ in the definition of V^h. Let U be the grid function defined in Section 3.2.4 and K be an interior simplex. Denote the characteristic function on K by χ_K. Assume the underlying mesh is a uniform Cartesian partition. Then, by

(3.19), we have

$$-\beta\cdot\left(q_h^+-q_h^-,\chi_K\right)_{\mathcal{T}_h}=-\sum_{i=1}^{d}\beta_i\left(\delta^+_{x_i,h_i}U_{\alpha_K}-\delta^-_{x_i,h_i}U_{\alpha_K}\right)=\sum_{i=1}^{d}\beta_i h_i\delta^2_{x_i,h_i}U_{\alpha_K}. \quad (3.27)$$

Also, by (3.20), we have

$$\begin{aligned}\alpha:&\left(P^{++}_{i,j}-P^{+-}_{i,j}-P^{-+}_{i,j}+P^{--}_{i,j},\chi_K\right)_{\mathcal{T}_h} \quad (3.28)\\ &=\sum_{i,j=1}^{d}\alpha_{i,j}\left(\delta^+_{x_i,h_i}\delta^+_{x_j,h_j}U_{\alpha_K}-\delta^+_{x_i,h_i}\delta^-_{x_j,h_j}U_{\alpha_K}-\delta^-_{x_i,h_i}\delta^+_{x_j,h_j}U_{\alpha_K}+\delta^-_{x_i,h_i}\delta^-_{x_j,h_j}U_{\alpha_K}\right)\\ &=\sum_{i,j=1}^{d}\alpha_{i,j}h_ih_j\delta^2_{x_i,h_i}\delta^2_{x_j,h_j}U_{\alpha_K}.\end{aligned}$$

Thus, for $\beta=\vec{1}$ and $\alpha=\mathbf{1}$, we again recover scaled approximations for the Laplace and biharmonic operator, as in Chapter 2.

We now suppose $r\geq 1$ in the definition of V^h. Let $i\in\{1,2,\ldots,d\}$. Observe, using the boundary conditions from Section 3.2.4, we have

$$\left(q_i^+-q_i^-,\phi\right)_{\mathcal{T}_h}=a_i^+\left(u_h,\phi\right)-a_i^-\left(u_h,\phi\right)=-\left\langle T_i^+(u_h)-T_i^-(u_h),\left[\phi\right]n_{e_i}\right\rangle_{\mathcal{E}_h^I}.$$

Thus,

$$-\beta\cdot\left(q_h^+-q_h^-,\phi\right)_{\mathcal{T}_h}=\sum_{i=1}^{d}\beta_i\left\langle T_i^+(u_h)-T_i^-(u_h),\left[\phi\right]n_{e_i}\right\rangle_{\mathcal{E}_h^I},$$

and with the correct labeling of the mesh, we have

$$-\beta\cdot\left(q_h^+-q_h^-,\phi\right)_{\mathcal{T}_h}=\sum_{i=1}^{d}\beta_i\left\langle\left[u_h\right],\left[\phi\right]n_{e_i}\right\rangle_{\mathcal{E}_h^I}. \quad (3.29)$$

Similarly, for $i, j \in \{1, 2, \ldots, d\}$,

$$\begin{aligned}
&\left(P_{i,j}^{++} - P_{i,j}^{+-} - P_{i,j}^{-+} + P_{i,j}^{--}, \phi\right)_{\mathcal{T}_h} \\
&\quad = a_j^+\left(q_i^+, \phi\right) - a_j^-\left(q_i^+, \phi\right) - a_j^+\left(q_i^-, \phi\right) + a_j^-\left(q_i^-, \phi\right) \\
&\quad = -\left\langle T_j^+(q_i^+) - T_j^-(q_i^+), \left[\phi\right] n_{e_j}\right\rangle_{\mathcal{E}_h^I} - \left\langle T_j^-(q_i^-) - T_j^+(q_i^-), \left[\phi\right] n_{e_j}\right\rangle_{\mathcal{E}_h^I}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\alpha : &\left(P_{i,j}^{++} - P_{i,j}^{+-} - P_{i,j}^{-+} + P_{i,j}^{--}, \phi\right)_{\mathcal{T}_h} \\
&= -\sum_{i,j=1}^{d} \alpha_{i,j}\left\langle T_j^+(q_i^+ - q_i^-) - T_j^-(q_i^+ - q_i^-), \left[\phi\right] n_{e_j}\right\rangle_{\mathcal{E}_h^I},
\end{aligned}$$

and with the same labeling of the mesh, we have

$$\alpha : \left(P_{i,j}^{++} - P_{i,j}^{+-} - P_{i,j}^{-+} + P_{i,j}^{--}, \phi\right)_{\mathcal{T}_h} = \sum_{i,j=1}^{d} \alpha_{i,j}\left\langle \left[q_i^- - q_i^+\right], \left[\phi\right] n_{e_j}\right\rangle_{\mathcal{E}_h^I}. \tag{3.30}$$

From above, we can see that

$$\begin{aligned}
&a_0\big(u_h, q_h^+, q_h^-, P_h^{++}, P_h^{+-}, P_h^{-+}, P_h^{--}; \phi_h\big) \\
&\quad = \left(F\left(P_h, q_h, u_h, \cdot\right), \phi_h\right)_{\mathcal{T}_h} + \sum_{i=1}^{d} \beta_i\left\langle \left[u_h\right], \left[\phi_h\right] n_{e_i}\right\rangle_{\mathcal{E}_h^I} \\
&\qquad + \sum_{i,j=1}^{d} \alpha_{i,j}\left\langle \left[q_i^- - q_i^+\right], \left[\phi_h\right] n_{e_j}\right\rangle_{\mathcal{E}_h^I}
\end{aligned} \tag{3.31}$$

where

$$P_h = \frac{P_h^{+-} + P_h^{-+}}{2}, \qquad q_h = \frac{q_h^+ + q_h^-}{2},$$

and q_h^+, q_h^- are both approximations for ∇u. Thus, adding a numerical moment and a numerical viscosity amounts to the addition of interior jump/stabilization terms to an L^2-projection of the fully nonlinear PDE operator into V^h. We do note that

the jump/stabilization terms that arise due to the numerical moment penalize the differences in q_h^+ and q_h^-. Thus, the numerical moment is not analogous to a high order penalization term that penalizes jumps in a single approximation for ∇u. Instead, it penalizes the difference in two optimal DG approximations for ∇u (cf. [26]).

3.2.7 Remarks about the Formulation

We conclude this section with a few remarks and a convergence theorem.

Remark 3.3.

(a) Looking backwards, (3.16) provides the proper interpretation for each of q_h^μ and $P_h^{\mu\nu}$, $\mu, \nu = -, +$, for a given function u_h. Each $q_h^\pm$ defines a discrete gradient for u_h, and each $P_h^{\mu\nu}$ defines a discrete Hessian for u_h. The functions q_h^- and q_h^+ should be very close to each other if ∇u exists and is continuous. Similarly, the functions P_h^{--}, P_h^{-+}, P_h^{+-}, and P_h^{++} should be very close to each other if D^2u exists and is continuous. However, their discrepancies are expected to be large if ∇u or D^2u, respectively, does not exist or is not continuous. The auxiliary functions $q_h^\pm$ defined by (3.16a) and the auxiliary functions P_h^{++}, P_h^{+-}, P_h^{-+}, and P_h^{--} defined by (3.16b) can be regarded as high order extensions of their lower order counterparts defined in Chapter 2 using a FD methodology.

(b) We saw that the linear equations defined by (3.16) are linearly independent given a value for the trace of $\frac{P_h^{+-}+P_h^{-+}}{2}$. In fact, there is a symmetric, positive definite mapping from the trace of $\frac{P_h^{+-}+P_h^{-+}}{2}$ to u_h. The proof for this fact is given in [40].

(c) Notice that (3.12) and (3.16) form a nonlinear system of equations where the nonlinearity only appears in a_0. Thus, a nonlinear solver is necessary in implementing the above scheme. In Section 3.5, an iterative method is used with a "trivial" initial guess for the numerical tests. Since a good initial guess is essential for most nonlinear solvers to converge, another possibility is to

first linearize the nonlinear operator and solve the resulting linear system first. However, we show in our numerical tests that the "trivial" initial guess works well in many cases. We suspect that the g-monotonicity of $\widehat{F}$ provided by using a numerical operator enlarges the domain of "good" initial guesses over which the iterative method converges. A few more comments about solvers will be discussed in Section 3.4.

From the proposed boundary conditions, we can see that the relationship $P_h^{-+} = P_h^{+-}$ has been extended to the boundary when using a uniform Cartesian mesh. Furthermore, using the function extensions defined above to define ghost values and substituting equations (3.19) and (3.20) into (3.12), we successfully recover the convergent FD method defined in Chapter 2 for the grid function U when using $r = 0$ on a uniform Cartesian mesh. Thus, we have the following convergence result for the LDG methods formulated in this section:

Definition 3.1. *Suppose the LDG method given by (3.12) and (3.16) has a solution. A solution $(u_h, q_h^+, q_h^-, P_h^{++}, P_h^{+-}, P_h^{-+}, P_h^{--})$ is said to be* stable *if there exists a constant $C > 0$, which is independent of h, such that u_h satisfies*

$$\|u_h\|_{\ell^\infty(\Omega)} \leq C.$$

Also, the LDG method is said to be a stable scheme *if all of its solutions are stable solutions.*

Theorem 3.1. *Let Ω be a d-rectangle. Suppose problem (3.1) satisfies the comparison principle of Definition 1.4, has a unique continuous viscosity solution u, and the operator F is independent of $u_{x_i x_j}$ for all $i, j \in \{1, 2, \ldots, d\}$ such that $i \neq j$. Also, suppose $\widehat{F}$ in (3.12) depends only on $P_h^{++} + P_h^{--}$, $P_h^{+-} + P_h^{-+}$, q_h^+, q_h^-, u_h, and x. Let $(u_h, q_h^+, q_h^-, P_h^{++}, P_h^{+-}, P_h^{-+}, P_h^{--})$ be a solution to a consistent, g-monotone, and stable LDG method (3.12) and (3.16). Then, for $r = 0$ and $\mathcal{T}_h$ a uniform Cartesian mesh, u_h converges to u locally uniformly as $h \to 0^+$.*

Remark 3.4. *The convergence theory of the proposed LDG method is currently limited to $\mathcal{T}_h$ a uniform Cartesian mesh with $r = 0$. Heuristically, using higher order elements should increase the rate and/or accuracy of convergence, as will be seen in the numerical tests provided in Section 3.5.*

3.3 Extensions of the LDG Framework to Second Order Parabolic Problems

We now develop fully discrete methods for approximating the initial-boundary value problem (3.2) using an LDG spatial-discretization paired with the method of lines for the time discretization. Taking advantage of the elliptic formulation in Section 3.2, we will propose the following implicit and explicit time-discretizations: forward Euler, backward Euler, trapezoidal, and Runge-Kutta (RK). The time-discretization used in application should be dictated by the potential optimal order of the LDG spatial-discretization which is given by $r + 1$ for sufficiently regular viscosity solutions. The proposed methods will be tested in Section 3.5.4.

We first develop the semi-discrete discretization of the (fully) nonlinear equation (3.2a) by discretizing the spatial dimension. Replacing the PDE operator F with a numerical operator $\widehat{F}$ in (3.2a), applying a spatial discretization using the above LDG framework for elliptic equations, and using the projection operator $\mathcal{P}_h$ defined by (3.4), we have the following semi-discrete equation

$$u_{h_t} = -\mathcal{P}_h\Big(\widehat{F}\left(P_h^{++}, P_h^{+-}, P_h^{-+}, P_h^{--}, q_h^+, q_h^-, u_h, x, t\right)\Big), \tag{3.32}$$

where, given u_h at time t, corresponding values for $q_h^{\pm}$ and $P_h^{\mu\nu}$, $\mu, \nu \in \{+, -\}$, can be found using the methodology below.

We now develop the full-discretization of (3.2a) by applying an ODE solver to the semi-discrete (variational) form given in (3.32). To partition the time domain, we fix an integer $M > 0$ and let $\Delta t = \frac{T}{M}$. Then, we define $t_k := k\,\Delta t$ for a real

number k with $0 \leq k \leq M$. Notationally, $u_h^k(x) \in V^h$ and $q_h^{\pm,k} \in \left[V^h\right]^d$ will be an approximation for $u(x,t_k)$ and $q_h^{\pm}(x,t_k)$, respectively, for all $0 \leq k \leq M$. For both implicit and explicit schemes, we define the initial value, u_h^0, by

$$u_h^0 = \mathcal{P}_h u_0. \tag{3.33}$$

To simplify the appearance of the methods and to make them more transparent for use with a given ODE solver, we define discrete "one-sided" gradient operators $\nabla_{h,k}^{\pm} : V_h \to \left[V^h\right]^d$ at time t_k using (3.16a) and discrete "one-sided" Hessian operators $D_{h,k}^{\mu\nu} : V_h \to \left[V^h\right]^{d\times d}$ for $\mu,\nu \in \{+,-\}$ at time t_k using (3.16b), where $0 \leq k \leq M$. Then, we have

$$\begin{aligned}\left(\left[\nabla_{h,k}^{\pm} v\right]_i, \phi_i^{\pm}\right)_{\mathcal{T}_h} &= \left\langle T_i^{\pm}(v), \left[\phi_i^{\pm}\right] n_{e_i}\right\rangle_{\mathcal{E}_h^I} + \left\langle T_i^{\pm}(v), \phi_i^{\pm}(x^I)\, n_i\right\rangle_{\mathcal{E}_h^B} \\ &\quad - \left(v, (\phi_i^{\pm})_{x_i}\right)_{\mathcal{T}_h} \qquad \forall \phi_i^{\pm} \in V^h\end{aligned} \tag{3.34}$$

for $i = 1, 2, \ldots, d$, for all $v \in V^h$, where we also enforce the time-dependent constraints given by either

$$\sum_{i=1}^{d} \left\langle v, \varphi_h\, n_i\right\rangle_{\mathcal{E}_h^B} = \sum_{i=1}^{d} \left\langle g(\cdot,t_k), \varphi_h\, n_i\right\rangle_{\tilde{\mathcal{E}}_h^B} \qquad \forall \varphi_h \in V^h$$

by using (3.18) when $r \geq 1$ or

$$\sum_{i=1}^{d} \left\langle v(x^E), n_i\right\rangle_e = \sum_{i=1}^{d} \left\langle g(\cdot,t_k), n_i\right\rangle_{\tilde{e}}$$

for each $e \in \mathcal{E}_h^B$ by combining (3.21) and (3.22) when $r = 0$. Also, we have

$$\begin{aligned}\left(\left[D_{h,k}^{\mu\nu} v\right]_{i,j}, \psi_{i,j}^{\mu\nu}\right)_{\mathcal{T}_h} &= \left\langle T_j^{\nu}\left(q_{i,k}^{\mu}\right), \left[\psi_{i,j}^{\mu\nu}\right] n_{e_j}\right\rangle_{\mathcal{E}_h^I} + \left\langle T_j^{\nu}(q_{i,k}^{\mu}), \psi_{i,j}^{\mu\nu}(x^I)\, n_j\right\rangle_{\mathcal{E}_h^B} \\ &\quad - \left(q_{i,k}^{\mu}, (\psi_{i,j}^{\mu\nu})_{x_j}\right)_{\mathcal{T}_h} \qquad \forall \psi_{i,j}^{\mu\nu} \in V^h\end{aligned} \tag{3.35}$$

for $q^{\mu}_{i,k} := \left(\nabla^{\mu}_{h,k} v\right)_i$, for $i, j \in \{1, 2, \ldots, d\}$, $\mu, \nu \in \{+, -\}$, and for all $v \in V^h$, where we assume $q^{\pm}_{i,k}(x^E) = q^{\pm}_{i,k}(x)$ when $r \geq 1$ or

$$\sum_{i=1}^{d} \left\langle q^{\pm}_{i,k}\left(x^I\right) - q^{\pm}_{i,k}\left(x^E\right), n_{e_i} \right\rangle_e = 0$$

and

$$\begin{aligned} q^{-}_{i,k}(x^E) &= q^{+}_{i,k}(x^I), \qquad \text{if } n_{e_i} < 0, \\ q^{+}_{i,k}(x^E) &= q^{-}_{i,k}(x^I), \qquad \text{if } n_{e_i} \geq 0 \end{aligned}$$

for all $e \in \mathcal{E}^B_h$, using (3.24) and (3.25), when $r = 0$. Note, for $k = 0$, we replace $g(\cdot, t_k)$ with $u_0(\cdot)$ in the above constraint equations.

To simplify the presentation of the fully-discrete methods, we introduce the operator notation

$$\widehat{F}^k[v] := \widehat{F}\left(D^{++}_{h,k} v, D^{+-}_{h,k} v, D^{-+}_{h,k} v, D^{--}_{h,k} v, \nabla^{+}_{h,k} v, \nabla^{-}_{h,k} v, v, x, k\,\Delta t\right) \tag{3.36}$$

for all $v \in V^h$. Then, the semi-discrete equation can be rewritten compactly as

$$\left(u_h\right)_t(x, t_k) = -\mathcal{P}_h \widehat{F}^k\left[u_h(x, t_k)\right] \tag{3.37}$$

for all $0 \leq k \leq M$, $x \in \Omega$.

Lastly, we define a modified projection operator $\mathcal{P}_{h,k} : L^2(\mathcal{T}_h) \to V^h$ that will be used to enforce the boundary conditions for explicit methods using a penalty technique due to Nitsche in [45]. Thus, we define $\mathcal{P}_{h,k}$ by

$$\begin{aligned} &\left(\mathcal{P}_{h,k} v, \varphi_h\right)_{\mathcal{T}_h} + \delta \sum_{i=1}^{d} \left\langle \mathcal{P}_{h,k} v, \varphi_h\, n_i \right\rangle_{\mathcal{E}^B_h} \\ &\quad = \left(v, \varphi_h\right)_{\mathcal{T}_h} + \delta \sum_{i=1}^{d} \left\langle g(\cdot, t_k), \varphi_h\, n_i \right\rangle_{\widetilde{\mathcal{E}}^B_h} \qquad \forall \varphi_h \in V^h \end{aligned} \tag{3.38}$$

for all $v \in L^2(\mathcal{T}_h)$, where δ is a nonnegative penalty constant and $0 \leq k \leq M$. We note that, for $\delta = 0$, $\mathcal{P}_{h,k} = \mathcal{P}_h$, yielding the broken L^2-projection operator.

Using the above conventions, we can define fully discrete methods for approximating problem (3.2) based on approximating (3.37) using the forward Euler method, backward Euler method, or the trapezoidal method. Hence, we have the following fully discrete schemes for approximating (3.2):

$$u_h^{n+1} = \mathcal{P}_{h,n+1}\left(u_h^n - \Delta t\, \widehat{F}^n\left[u_h^n\right]\right), \tag{3.39}$$

$$u_h^{n+1} + \Delta t\, \mathcal{P}_h\, \widehat{F}^{n+1}\left[u_h^{n+1}\right] = u_h^n, \tag{3.40}$$

and

$$u_h^{n+1} + \frac{\Delta t}{2}\, \mathcal{P}_h\, \widehat{F}^{n+1}\left[u_h^{n+1}\right] = u_h^n - \frac{\Delta t}{2}\, \mathcal{P}_h\, \widehat{F}^n\left[u_h^n\right] \tag{3.41}$$

for $n = 0, 1, \ldots, M-1$, where $u_h^0 := \mathcal{P}_h u_0$ and, for (3.40) and (3.41), we also have the implied auxiliary linear equations

$$\begin{aligned} q_h^{\mu,n} &= \nabla_{h,n}^{\mu} u_h^n \qquad \forall \mu \in \{+,-\}, \\ P_h^{\mu\nu,n} &= D_{h,n}^{\mu\nu} u_h^n \qquad \forall \mu, \nu \in \{+,-\} \end{aligned}$$

by (3.36). Observe, (3.39), (3.40), and (3.41) correspond to the forward Euler method, backward Euler method, and trapezoidal method, respectively.

Remark 3.5. *Using an implicit method such as the backward Euler method or the trapezoidal method is equivalent to approximating a fully nonlinear elliptic PDE at each time level $n = 1, 2, \ldots, M$ using the LDG methods for elliptic PDEs formulated in Section 3.2. Due to the time integration, the nonlinear solver has a natural initial guess for each time-step given by the approximation for u at the previous time step.*

We now consider using RK methods for approximating (3.37). Let s be a positive integer, $A \in \mathbb{R}^{s\times s}$, and $b, c \in \mathbb{R}^s$ such that

$$\sum_{\ell=1}^{s} a_{k,\ell} = c_k$$

for each $k = 1, 2, \ldots, s$. Then, a generic s-stage RK method for approximating (3.37) is defined by

$$u_h^{n+1} = \mathcal{P}_{h,n+1}\left(u_h^n - \Delta t \sum_{\ell=1}^{s} b_\ell \widehat{F}^{n+c_\ell}[\xi_h^{n,\ell}]\right) \tag{3.42}$$

with

$$\xi_h^{n,\ell} = \mathcal{P}_{h,n+c_k}\left(u_h^n - \Delta t \sum_{k=1}^{s} a_{k,\ell} \widehat{F}^{n+c_k}[\xi_h^{n,k}]\right)$$

for all $n = 0, 1, \ldots, N-1$ and $u_h^0 = \mathcal{P}_h u_0$. We note that (3.42) corresponds to an explicit method when A is strictly lower diagonal and an implicit method otherwise.

Remark 3.6.

(a) We can interpret $\xi_h^{n,\ell}$ in (3.42) as an approximation for $u_h^{n+c_\ell}$. Since the boundary condition at time t_{n+1} is enforced by $\widehat{F}^{n+1}$, we can set $\delta = 0$ in (3.38) if $c_s = 1$.

(b) We do not analyze the CFL condition required for the above explicit schemes. In Section 3.5.4, we implement the above methods and record the observed CFL conditions.

3.4 General Solvers

We now discuss different strategies for solving the nonlinear system of equations that results from the proposed LDG discretization for the elliptic problem. The underlying goal for the methodology presented in this chapter is to discretize the fully nonlinear PDE problem in a way that removes much of the burden of approximating viscosity solutions from the design of the solver. Thus, our primary focus is at the discretization

level. However, some of the properties of the methodology are more apparent from the solver perspective.

Most tests show that it is sufficient to simply use a Newton solver on the full system of equations given in the above mixed formulation, (3.12) and (3.16). Observe, only (3.12) is nonlinear, the equation is purely algebraic, and $\widehat{F}$ is monotone in six of its arguments. The auxiliary equations (3.16) are all linear. The numerical operator presented in this paper is symmetric in both the mixed approximations P_h^{-+} and P_h^{+-} and the non-mixed approximations P_h^{--} and P_h^{++}. Thus, we can reduce the size of the system of equations by averaging the two pairs of auxiliary variables in the above formulation without changing the methodology. However, for some choices of the numerical operator, such a reduction of variables may not be possible.

Due to the size of the mixed formulation, we first provide a splitting algorithm that provides an alternative to straight-forward Newton solvers for the entire system of equations. By using a splitting algorithm, the resulting algorithm will iteratively solve an entirely local, nonlinear equation that has strong monotonicity properties in the d unknown arguments, and the solution of the equation can be mapped to an updated approximation for u_h. Tests show that the solver is particularly useful for nonlinear problems that have a unique viscosity solution only defined in a restrictive function class. However, the solver is not as efficient as the second solver we present that takes advantage of the above nonstandard discretization technique. In order to improve the speed of the solver, fast Poisson solvers for the proposed discretization need to be developed.

The second solver strategy that we present is a natural generalization of the FD methodology for numerical PDEs presented in Chapter 2. Careful examination of the auxiliary linear equations in the mixed formulation reveals that operators can be statically computed that map u_h to each auxiliary variable at the cost of sparse matrix multiplication and addition as well as inverting the local mass matrices. Thus, all auxiliary equations in the mixed formulation can be solved for a given function u_h.

Substituting these operators directly into the numerical operator results in a single nonlinear variational problem for u_h.

3.4.1 An Inverse-Poisson Fixed-Point Solver

We now propose the above mentioned splitting algorithm that takes into account the special structure of the nonlinear algebraic system that results from the above LDG discretization methods for elliptic PDEs and parabolic PDEs when using implicit time-stepping. The algorithm is strongly based upon using a particular numerical moment and the results of Section 3.2.5.

Algorithm 3.1.

1. *Pick an initial guess for* u_h.

2. *Form initial guesses for* q_h^+, q_h^-, P_h^{++}, P_h^{+-}, P_h^{-+}, *and* P_h^{--} *using equations* (3.16).

3. *Set*

$$[G]_i := F\left(\frac{P_h^{-+}+P_h^{+-}}{2}, \frac{q_h^-+q_h^+}{2}, u_h, x\right) + \gamma\left[P_h^{++} - P_h^{+-} - P_h^{-+} + P_h^{--}\right]_{i,i} - \beta_i\left[q_h^+ - q_h^-\right]_i$$

   for a fixed constant $\gamma > 0$, *and solve*

$$\Big([G]_i, \varphi_i\Big)_{\mathcal{T}_h} = 0 \qquad \forall \varphi_i \in V^h$$

   for $\left[\frac{P_h^{-+}+P_h^{+-}}{2}\right]_{i,i}$ *for all* $i = 1, 2, \ldots, d$. *For sufficiently large* γ *and a differentiable operator* F, *the above set of equations has a negative definite Jacobian.*

4. *Find* u_h, q_h^+, *and* q_h^- *by solving the linear system of equations formed by* (3.16a) *and the trace of averaging* (3.16b) *for* $\mu = -, \nu = +$ *and* $\mu = +, \nu = -$.

Observe, this is equivalent to solving Poisson's equation with source data given by the trace of $\frac{P_h^{-+}+P_h^{+-}}{2}$. *Another method that yields* u_h *directly can be found in [26], where again the trace of* $\frac{P_h^{-+}+P_h^{+-}}{2}$ *is used as the source data for Poisson's equation.*

5. *Solve* (3.16b) *for* P_h^{++}, P_h^{+-}, P_h^{-+}, *and* P_h^{--}. *If the alternative approach in step 4 was used, also solve* (3.16a) *for* q_h^+ *and* q_h^-.

6. *Repeat Steps 3 - 5 until the change in* $\frac{P_h^{-+}+P_h^{+-}}{2}$ *is sufficiently small.*

We now make a couple of comments about the proposed solver.

Remark 3.7.

(a) The proposed algorithm is well-posed. By the g-monotonicity assumption, the nonlinear equation in Step 3 has a root. By the lemmas in Section 3.2.5, inverting the linear system in Step 4 is possible. We also note that the nonlinear equation in Step 3 is entirely local with respect to the unknown variable.

(b) Clearly a fixed point for the solver corresponds to a discrete solution of the original PDE problem. In Sections 3.5.1, 3.5.2, and 3.5.3, we demonstrate that the above solver can be used to eliminate numerical artifacts that arise due to low-regularity PDE artifacts, as discussed in Section 1.3. Thus, the proposed solver is less dependent upon the initial guess. The algorithm can also be used to form a preconditioned initial guess for other nonlinear solvers that may be faster but require a "better" initial guess.

3.4.2 A Direct Approach for a Reduced System

In this section, we propose a solver technique that is analogous to the approach used in the FD methodology of Chapter 2. We first build numerical gradient and Hessian operators ∇_h^μ and $D_h^{\mu\nu}$ that act on a function $v_h \in V^h$ by using the operators defined in (3.34) and (3.35) with the time dependence dropped from the definition. Observe,

if $\left(u_h, q_h^+, q_h^-, P_h^{++}, P_h^{+-}, P_h^{-+}, P_h^{--}\right)$ is a solution to (3.12) and (3.16), then $q_h^\mu = \nabla_h^\mu u_h$ and $P_h^{\mu\nu} = D_h^{\mu\nu} u_h$ for all $\mu, \nu \in \{+, -\}$. Furthermore, the operators ∇_h^μ and $D_h^{\mu\nu}$ can be statically computed with the cost of inverting the local mass matrix and sparse matrix addition and multiplication.

Using these numerical derivative operators, the second solver is given by:

Algorithm 3.2.

1. *Given $\mathcal{T}_h$ and V^h, compute the operators ∇_h^μ and $D_h^{\mu\nu}$.*

2. *Solve the single nonlinear equation*

$$0 = \left(\widehat{F}\left(D_h^{++}u_h, D_h^{+-}u_h, D_h^{-+}u_h, D_h^{--}u_h, \nabla_h^+ u_h, \nabla_h^- u_h, u_h, \cdot\right), \varphi_h\right)_{\mathcal{T}_h}$$

   for all $\varphi_h \in V^h$.

We note that a reduced formulation can also be used where we simply create the new discrete Hessian operators

$$\overline{D}_h^2 := \frac{D_h^{--} + D_h^{++}}{2}, \qquad \widetilde{D}_h^2 := \frac{D_h^{-+} + D_h^{+-}}{2}.$$

In this case, the Lax-Friedrichs-like numerical operator becomes

$$\begin{aligned} &\widehat{F}\left(\overline{D}_h^2 u_h, \widetilde{D}_h^2 u_h, \nabla_h^+ u_h, \nabla_h^- u_h, u_h, x\right) \\ &\quad = F\left(\widetilde{D}_h^2 u_h, \frac{\left(\nabla_h^- + \nabla_h^+\right) u_h}{2}, u_h, x\right) + 2\alpha : \left(\overline{D}_h^2 u_h - \widetilde{D}_h^2 u_h\right) - \beta \cdot \left(\nabla_h^+ u_h - \nabla_h^- u_h\right). \end{aligned} \tag{3.43}$$

For all of the tests below where a Newton solver is used for the full system of equations in the mixed formulation, analogous results were obtained using Algorithm 3.2 with the reduced numerical operator. As expected, for two-dimensional problems we observed significant speed-up in the performance of the solver.

Remark 3.8. *The methodology of Algorithm 3.2 follows directly from the FD methodology where derivatives in a PDE are simply replaced by numerical derivatives of the approximation for the solution u to form the discretization of the PDE problem. For nonlinear problems, we replace the nonlinear PDE operator by a numerical operator. In our LDG setting, we use LDG methodologies to define the various numerical derivatives.*

3.5 Numerical Experiments

We now present a series of numerical tests to demonstrate the utility of the proposed LDG methods for fully nonlinear PDE problems of type (3.1) and (3.2). For elliptic problems, both Monge-Ampère and Hamilton-Jacobi-Bellman types of equations will be tested. The one-dimensional tests use spatial meshes composed of uniform intervals, and the two-dimensional tests use spatial meshes composed of uniform rectangles. To solve the resulting nonlinear algebraic systems, we use either the Matlab built-in nonlinear solver *fsolve* or Algorithm 3.1, where *fsolve* is used to perform Step 3 of Algorithm 3.1. For elliptic problems in one-dimension, we choose the initial guess as $\overline{u}$, the secant line formed by the boundary data. For elliptic problems in two-dimensions, we choose the initial guess as the zero function. In contrast, we choose the initial guess as the approximation formed at the previous time step for implicit discretizations of parabolic problems. We also choose the approximation at time $t = 0$ to be given by the L^2-projection of the initial condition into V^h. The role of the numerical moment will be further explored in Section 3.5.3.

For our numerical tests, errors will be measured in the L^∞ norm and the L^2 norm. For elliptic problems and parabolic problems where the error is not dominated by the time discretization, it appears that the spatial errors are of order $\mathcal{O}(h^s)$ for most problems, where $s = \min\{r+1, k\}$ for the viscosity solution $u \in H^k(\Omega)$. Thus, the schemes appear to exhibit an optimal rate of convergence in both norms (cf. [9]). However, for a couple of problems, we do observe less than optimal rates of

convergence. We note that the actual convergence rates have not yet been analyzed, and they may also depend on the regularity of the differential operator F in addition to the regularity of the viscosity solution u.

3.5.1 One-Dimensional Elliptic Problems

We first present the results for four test problems of type (3.1) in one-dimension.

Example 3.1. *Consider the Monge-Ampère problem*

$$\begin{aligned} -u_{xx}^2 + 1 = 0, \qquad & 0 < x < 1, \\ u(0) = 0, \quad u(1) = 1/2. & \end{aligned}$$

This problem has exactly two classical solutions

$$u^+(x) = \frac{1}{2}x^2, \qquad u^-(x) = -\frac{1}{2}x^2 + x,$$

where u^+ is convex and u^- is concave. However, u^+ is the unique viscosity solution.

We approximate the given problem for various degree elements ($r = 0, 1, 2$) to see how the approximation converges with respect to h. Note, when $r = 0, 1$, the solution is not in the DG space V^h. The numerical results are shown in Table 3.1 and Figure 3.1. We observe that the approximations using $r = 2$ are almost exact for each mesh size. This is consistent with the fact $u^+ \in V^h$ when $r = 2$. In Section 3.5.3 we shall give additional insights about the selectiveness of our schemes.

Example 3.2. *Consider the problem*

$$\begin{aligned} -u_{xx}^3 + u_{xx} + S(x)^3 - S(x) = 0, \qquad & -1 < x < 1, \\ u(-1) = -\sin(1) - 8\cos(0.5) + 9, \quad u(1) = \sin(1) - 8\cos(0.5) + 9, & \end{aligned}$$

Table 3.1: Rates of convergence for Example 3.1 using $\alpha = 10$ and *fsolve* with initial guess $\mathcal{P}_h\overline{u}$.

r	Norm	$h = 1/4$	$h = 1/8$		$h = 1/16$		$h = 1/32$	
		Error	Error	Order	Error	Order	Error	Order
0	L^2	7.1e-02	3.5e-02	1.02	1.4e-02	1.30	7.5e-03	0.92
	L^∞	1.3e-01	8.7e-02	0.57	5.3e-02	0.73	2.9e-02	0.87
1	L^2	1.6e-02	5.0e-03	1.67	1.3e-03	1.90	3.4e-04	1.95
	L^∞	2.2e-02	6.3e-03	1.84	1.6e-03	2.00	3.9e-04	2.00
2	L^2	3.1e-13	3.0e-13	0.03	3.0e-13	-0.01	3.1e-13	-0.01
	L^∞	7.4e-13	6.1e-13	0.28	6.7e-13	-0.14	7.1e-13	-0.08

where

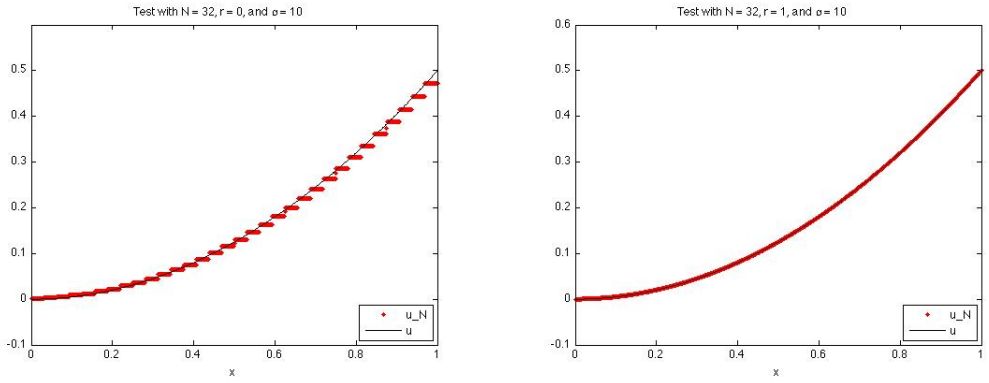
$$S(x) = \begin{cases} \frac{2x}{|x|}\cos(x^2) - 4x^2\sin(x|x|) + 2\cos(\frac{x}{2}) + 2, & x \neq 0, \\ 4, & x = 0. \end{cases}$$

This problem has the exact solution $u(x) = \sin(x|x|) - 8\cos\left(\frac{x}{2}\right) + x^2 + 8 \in H^2(-1,1)$.

Note that this problem is not monotone decreasing in u_{xx}, and the exact solution is not twice differentiable at $x = 0$. However, the derivative of F with respect to u_{xx} is uniformly bounded. The numerical results are shown in Table 3.2 and Figure 3.2. As expected, we can see from the plot that the error appears largest around the point $x = 0$, and both the accuracy and order of convergence improve as the order of the elements increases. For finer meshes, we see the rates of convergence begin to deteriorate. Theoretically, we expect the convergence rates to be bounded by two for high-order bases due to the lower regularity of the solution. We do note that the point $x = 0$ is a node for the partition.

Example 3.3. *Consider the stationary Hamilton-Jacobi-Bellman problem over a finite dimensional set*

$$\min_{\theta(x)\in\{1,2\}} \{-\theta u_{xx} + u_x - u + S(x)\} = 0, \qquad -1 < x < 1,$$
$$u(-1) = -1, \quad u(1) = 1,$$



(a) $r = 0$. **(b)** $r = 1$.

Figure 3.1: Computed solutions for Example 3.1 using $h =$ 3.125e-02, $\alpha = 10$, and *fsolve* with initial guess $\mathcal{P}_h\overline{u}$.

where

$$S(x) = \begin{cases} -12x^2 - 4|x|^3 + x|x|^3, & x < 0 \\ 24x^2 - 4|x|^3 + x|x|^3, & x \geq 0. \end{cases}$$

This problem has the exact solution $u(x) = x|x|^3 \in H^4(-1,1)$ *corresponding to* $\theta(x) = 1$ *for* $x < 0$ *and* $\theta(x) = 2$ *for* $x \geq 0$.

Approximating the problem using various order elements, we have the following results recorded in Table 3.3 and Figure 3.3. Due to the regularity of the solution, we expect the rates of convergence to be bounded by four for high-order bases. We observe that the rates of convergence for $r = 0, 1, 2$ appear to be optimal on average, while the rates of convergence for $r = 3$ appear to be limited to three. However, we still see increased accuracy for $r = 3$ when compared to $r = 2$.

Example 3.4. *Consider the stationary Hamilton-Jacobi-Bellman problem*

$$\inf_{0 \leq \theta(x) \leq 1} \left\{ -\theta u_{xx} + \theta^2 x^2 u_x + \frac{1}{x} u + S(x) \right\} = 0, \qquad 1.2 < x < 4,$$
$$u(1.2) = 1.44 \ln 1.2, \quad u(4) = 16 \ln 4,$$

Table 3.2: Rates of convergence for Example 3.2 using $\alpha = 6$ and *fsolve* with initial guess $\mathcal{P}_h\overline{u}$.

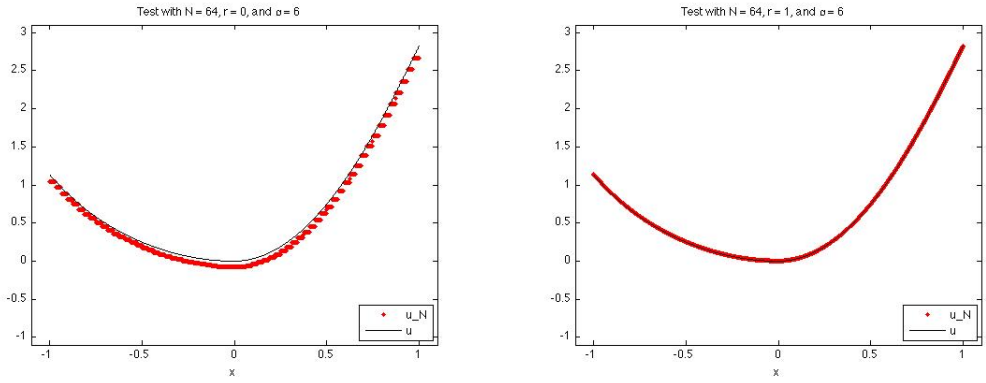
r	Norm	$h = 1/2$	$h = 1/4$		$h = 1/8$		$h = 1/16$		$h = 1/32$	
		Error	Error	Order	Error	Order	Error	Order	Error	Order
0	L^2	1.8	9.0e-01	1.04	4.3e-01	1.05	2.1e-01	1.04	1.0e-01	1.02
	L^∞	2.5	1.2	0.99	6.2e-01	1.00	3.1e-01	1.00	1.6e-01	1.00
1	L^2	2.9e-01	6.3e-02	2.20	1.9e-02	1.73	7.0e-03	1.44	2.8e-03	1.34
	L^∞	2.9e-01	6.4e-02	2.17	2.0e-02	1.66	7.6e-03	1.42	3.1e-03	1.30
2	L^2	5.7e-03	8.2e-04	2.80	1.3e-04	2.66	3.2e-05	2.03	9.1e-06	1.81
	L^∞	2.0e-02	3.1e-03	2.70	4.2e-04	2.87	5.5e-05	2.94	8.0e-06	2.77
3	L^2	8.8e-04	7.7e-05	3.51	3.0e-06	4.68	1.4e-07	4.42	1.0e-08	3.76
	L^∞	2.1e-03	1.4e-04	3.90	8.6e-06	4.01	5.6e-07	3.94	9.5e-08	2.57

where

$$S(x) = \frac{4\ln(x)^2 + 12\ln(x) + 9 - 8x^4\ln(x)^2 - 4x^4\ln(x)}{4x^3\left[2\ln(x) + 1\right]}.$$

This problem has the exact solution $u(x) = x^2 \ln x$ *corresponding to* $\theta(x) = \frac{2\ln(x)+3}{2x^3[2\ln(x)+1]}$.

Approximating the problem using various order elements, we obtain the results recorded in Table 3.4 and Figure 3.4. We can see that the approximations appear to reach a maximal level of accuracy of about 5.0e-7 in both the L^2 and the L^∞ norm.

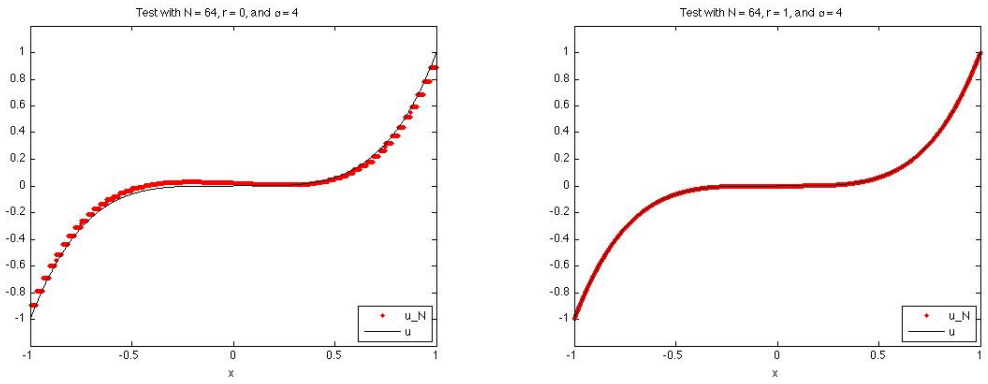


(a) $r = 0$. **(b)** $r = 1$.

Figure 3.2: Computed solutions for Example 3.2 using h = 3.125e-02, $\alpha = 6$, and *fsolve* with initial guess $\mathcal{P}_h\overline{u}$.

Table 3.3: Rates of convergence for Example 3.3 using $\alpha = 4$ and *fsolve* with initial guess $\mathcal{P}_h\overline{u}$.

r	Norm	$h = 1/2$	$h = 1/4$		$h = 1/8$		$h = 1/16$		$h = 1/32$	
		Error	Error	Order	Error	Order	Error	Order	Error	Order
0	L^2	5.0e-01	3.4e-01	0.53	1.3e-01	1.37	6.1e-02	1.12	4.4e-02	0.49
	L^∞	8.5e-01	5.7e-01	0.57	3.5e-01	0.72	2.0e-01	0.79	1.1e-01	0.86
1	L^2	1.4e-01	4.3e-02	1.72	9.7e-03	2.16	2.7e-03	1.85	7.3e-04	1.87
	L^∞	3.6e-01	1.1e-01	1.74	3.0e-02	1.87	7.8e-03	1.94	2.0e-03	1.97
2	L^2	2.8e-02	3.2e-03	3.10	4.0e-04	3.00	5.1e-05	2.99	6.4e-06	2.99
	L^∞	4.0e-02	5.5e-03	2.84	7.3e-04	2.93	9.3e-05	2.97	1.2e-05	2.98
3	L^2	9.4e-03	1.3e-03	2.91	1.6e-04	3.01	1.9e-05	3.01	2.4e-06	3.01
	L^∞	1.1e-02	1.5e-03	2.91	1.9e-04	3.02	2.3e-05	3.01	2.9e-06	3.01

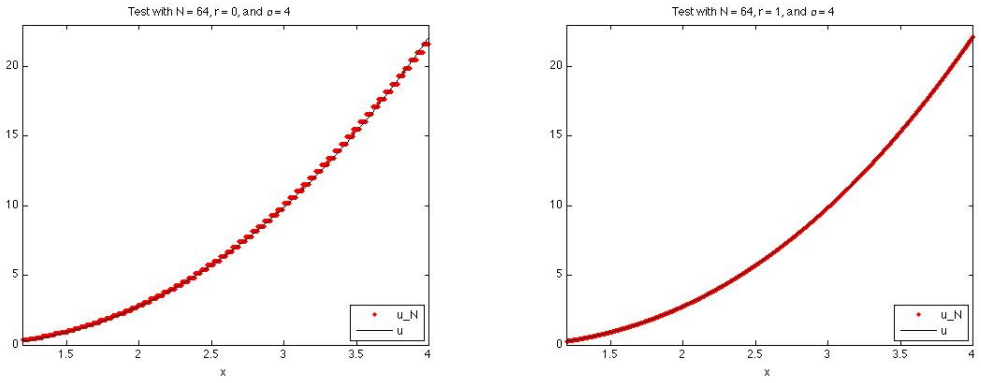


(a) $r = 0$. **(b)** $r = 1$.

Figure 3.3: Computed solutions for Example 3.3 using h = 3.125e-02, $\alpha = 4$, and *fsolve* with initial guess $\mathcal{P}_h\overline{u}$.

Table 3.4: Rates of convergence for Example 3.4 using $\alpha = 4$ and *fsolve* with initial guess $\mathcal{P}_h\overline{u}$.

r	Norm	$h = 2.8/4$	$h = 2.8/8$		$h = 2.8/16$		$h = 2.8/32$		$h = 2.8/64$	
		Error	Error	Order	Error	Order	Error	Order	Error	Order
0	L^2	5.2	3.3	0.67	1.5	1.08	6.1e-01	1.34	2.6e-01	1.25
	L^∞	5.7	3.6	0.65	1.9	0.91	1.1	0.84	5.7e-01	0.91
1	L^2	2.6e-01	8.6e-02	1.60	2.6e-02	1.72	7.4e-03	1.83	2.0e-03	1.90
	L^∞	3.3e-01	1.1e-01	1.56	3.5e-02	1.67	1.1e-02	1.71	3.4e-03	1.65
2	L^2	2.6e-03	3.9e-04	2.77	6.6e-05	2.55	1.4e-05	2.26	3.2e-06	2.09
	L^∞	7.3e-03	1.0e-03	2.85	1.4e-04	2.91	1.9e-05	2.81	4.1e-06	2.25
3	L^2	6.4e-05	4.2e-06	3.93	3.1e-07	3.75	1.2e-07	1.35	1.2e-07	0.08
	L^∞	2.7e-04	2.1e-05	3.72	1.4e-06	3.84	8.7e-07	0.72	8.8e-07	-0.01



(a) $r = 0$. **(b)** $r = 1$.

Figure 3.4: Computed solutions for Example 3.4 using h = 4.375e-02, $\alpha = 4$, and *fsolve* with initial guess $\mathcal{P}_h\overline{u}$.

3.5.2 Two-Dimensional Elliptic Problems

We now present the results for four test problems of type (3.1) in two-dimensions. Both Monge-Ampère and Hamilton-Jacobi-Bellman types of equations will be tested. We also perform a test using the infinite-Laplacian equation with a known low-regularity solution.

Example 3.5. *Consider the Monge-Ampère problem*

$$\begin{aligned} -\det D^2u = -u_{xx}\,u_{yy} + u_{xy}\,u_{yx} &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega, \end{aligned}$$

where $f = -(1 + x^2 + y^2)e^{x^2+y^2}$, $\Omega = (0,1) \times (0,1)$, *and* g *is chosen such that the viscosity solution is given by* $u(x,y) = e^{\frac{x^2+y^2}{2}}$.

Notice, the problem has two possible solutions as represented in Figure 3.5. Also, this problem is degenerate for the class of functions that are both concave and convex. Results for approximating with $r = 0, 1, 2$ can be found in Tables 3.5, 3.6, and 3.7,

respectively, where we observe optimal convergence rates. Plots for some of the various approximations can be found in Figures 3.6, 3.7, and 3.8.

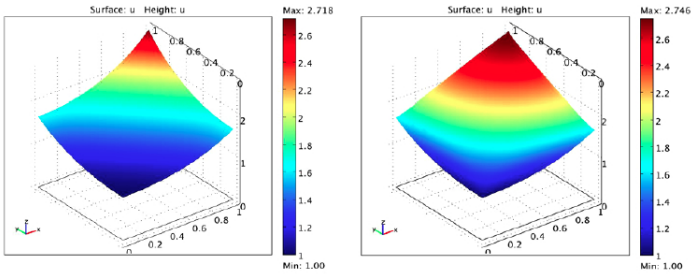


Figure 3.5: The two possible solutions for Example 3.5, as computed in [29]. The left plot corresponds to the viscosity solution while the right plot corresponds to the viscosity solution of $F[u] = \det D^2 u$.

Table 3.5: Rates of convergence for Example 3.5 using $r = 0$, $\alpha = 24I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h	L^∞ norm	order	L^2 norm	order
1.41e-01	3.73e-01		8.31e-02	
8.84e-02	2.42e-01	0.92	5.10e-02	1.04
5.89e-02	1.64e-01	0.95	3.31e-02	1.06
4.42e-02	1.24e-01	0.97	2.44e-02	1.07

Example 3.6. *Consider the Monge-Ampère problem*

$$\begin{aligned} -\det D^2 u = -u_{xx}\, u_{yy} + u_{xy}\, u_{yx} = 0 \qquad & \text{in } \Omega, \\ u = g \qquad & \text{on } \partial\Omega, \end{aligned}$$

where $\Omega = (-1,1) \times (-1,1)$ *and* g *is chosen such that the viscosity solution is given by* $u(x,y) = |x| \in H^1(\Omega)$.

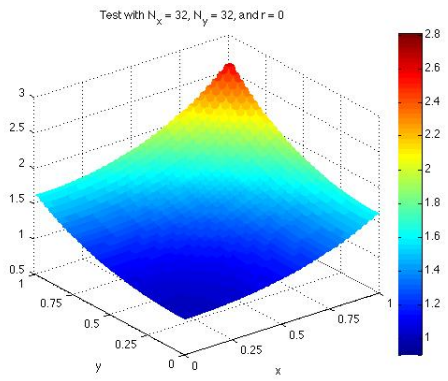


Figure 3.6: Computed solution for Example 3.5 using $r = 0$, $\alpha = 24I$, $h = 4.419\text{e-}02$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

Table 3.6: Rates of convergence for Example 3.5 using $r = 1$, $\alpha = 24I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h	L^∞ norm	order	L^2 norm	order
1.41e-01	2.47e-02		1.73e-03	
1.18e-01	1.36e-02	3.25	1.61e-03	0.39
1.01e-01	1.03e-02	1.81	1.12e-03	2.31
7.86e-02	8.04e-03	0.99	5.82e-04	2.62

Observe, the PDE is actually degenerate when acting on the solution u. Furthermore, due to the low regularity of u, we expect the rate of convergence to be bounded by one. Using both piecewise constant and piecewise linear basis functions, we can see that the rate of convergence is bounded by the theoretical bound in Table 3.8 and Table 3.9. Plots for some of the approximations can be found in Figure 3.8 for $r = 0$ and Figure 3.10 for $r = 1$. We remark that for $r = 0$, all three solver approaches discussed in Section 3.4 gave analogous results. However, for $r = 1$, the direct formulation has small residual wells that can trap the solver. Thus, for this test, the non-Newton solver given by Algorithm 3.1 appears to be better suited.

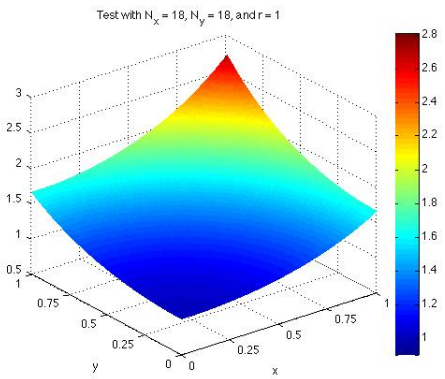


Figure 3.7: Computed solution for Example 3.5 using $r = 1$, $\alpha = 24I$, $h =$ 7.857e-02, and *fsolve* with initial guess $u_h^{(0)} = 0$.

Table 3.7: Rates of convergence for Example 3.5 using $r = 2$, $\alpha = 24I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h	L^∞ norm	order	L^2 norm	order
7.07e-01	6.39e-02		4.45e-03	
4.71e-01	2.32e-02	2.50	1.30e-03	3.03
3.54e-01	1.09e-02	2.63	5.45e-04	3.02

Example 3.7. *Consider the stationary Hamilton-Jacobi-Bellman problem*

$$\begin{aligned} \min\{-\Delta u, -\Delta u/2\} &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega, \end{aligned}$$

where $\Omega = (0, \pi) \times (-\pi/2, \pi/2)$,

$$f(x,y) = \begin{cases} 2\cos(x)\,\sin(y), & \text{if } (x,y) \in S, \\ \cos(x)\,\sin(y), & \text{otherwise,} \end{cases}$$

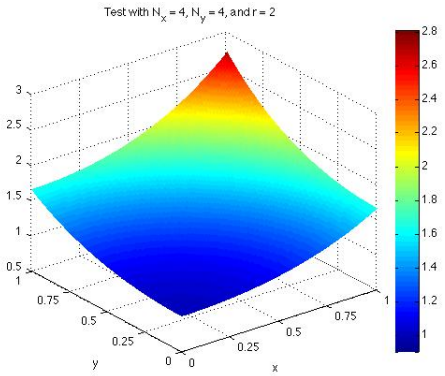


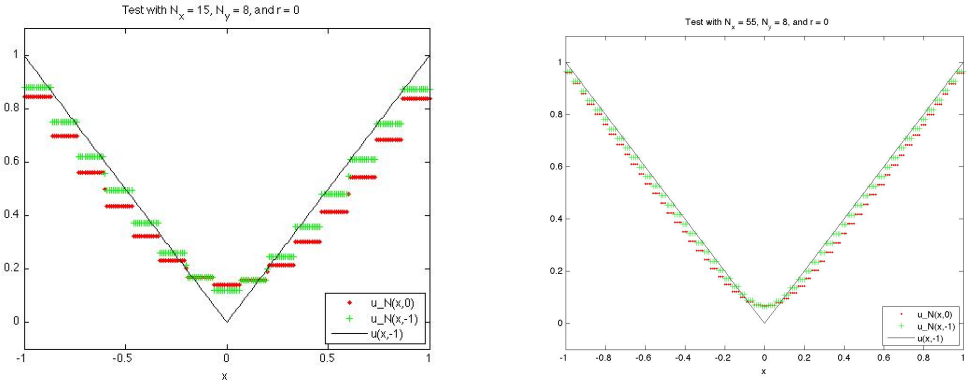
Figure 3.8: Computed solution for Example 3.5 using $r = 2$, $\alpha = 24I$, $h = 3.536\text{e-}01$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

Table 3.8: Rates of convergence for Example 3.6 using $r = 0$, $\alpha = I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h_x	L^∞ norm	order	L^2 norm	order
1.33e-01	1.87e-01		1.70e-01	
8.00e-02	1.30e-01	0.71	1.22e-01	0.65
5.71e-02	1.02e-01	0.72	9.77e-02	0.66
4.44e-02	8.51e-02	0.74	8.23e-02	0.68
3.64e-02	7.33e-02	0.74	7.16e-02	0.69

$S = (0, \pi/2] \times (-\pi/2, 0] \cup (\pi/2, \pi] \times (0, \pi/2)$, *and* g *is chosen such that the viscosity solution is given by* $u(x, y) = \cos(x)\,\sin(y)$.

We can see that the optimal coefficient for Δu varies over four patches in the domain. Results for approximating with $r = 0, 1, 2$ can be seen in Tables 3.10, 3.11, and 3.12, respectively, where we observe optimal convergence rates for $r = 0, 1$ and near optimal convergence rates for $r = 2$. Corresponding plots can be found in Figures 3.11, 3.12, and 3.13.



(a) $h_x = 6.667\text{e-}02$. **(b)** $h_x = 1.818\text{e-}02$.

Figure 3.9: Computed solutions for Example 3.6 using $r = 0$, $\alpha = I$, $h_y = 1.250\text{e-}01$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

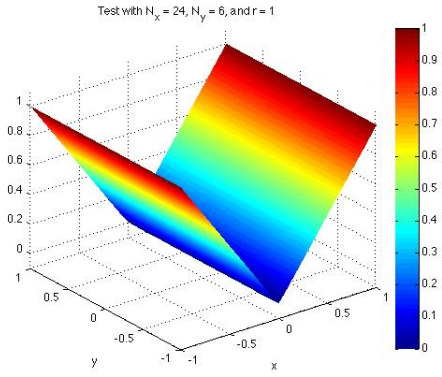
Table 3.9: Rates of convergence for Example 3.6 using $r = 1$, $\alpha = I$, $h_y = 1/3$ fixed, and Algorithm 3.1 with initial guess $u_h^{(0)} = 0$.

h_x	L^∞ norm	order	L^2 norm	order
2.50e-01	3.86e-02		3.42e-02	
1.25e-01	2.08e-02	0.89	1.85e-02	0.88
8.33e-02	1.38e-02	1.02	1.24e-02	0.99

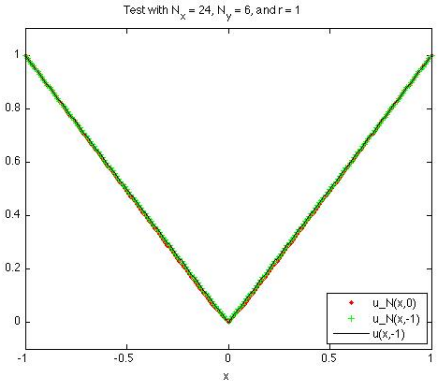
Example 3.8. *Consider the infinite-Laplacian problem*

$$\begin{aligned} -\Delta_\infty u := -u_{xx}\,u_x\,u_y - u_{xy}\,u_x\,u_y - u_{yx}\,u_y\,u_y - u_{yy}\,u_y\,u_y = 0 \qquad & \text{in } \Omega, \\ u = g \qquad & \text{on } \partial\Omega, \end{aligned}$$

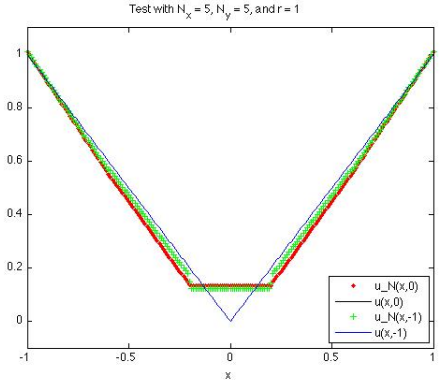
where $\Omega = (-1,1) \times (-1,1)$ *and* g *is chosen such that the viscosity solution is given by* $u(x,y) = |x|^{4/3} - |y|^{4/3}$. *While this problem is semilinear and not fully nonlinear, the solution has low regularity due to the fact* $u \in C^{1,\frac{1}{3}}(\overline{\Omega}) \cap H^1(\Omega)$.



(a) h_x = 4.167e-02 and h_y = 1.667e-01.



(b) h_x = 4.167e-02 and h_y = 1.667e-01.



(c) h_x = 2.000e-01 and h_y = 2.000e-01.

Figure 3.10: Computed solution for Example 3.6 using $r = 1$, $\alpha = I$, and Algorithm 3.1 with initial guess $u_h^{(0)} = 0$. Note, the top plots correspond to $x = 0$ an edge and the bottom plot does not.

By approximation theory, we expect the error to be bounded by $\mathcal{O}(h^1)$ independent of the degree of the polynomial basis. The approximation results for $r = 0, 1, 2$ can be found in Tables 3.13, 3.14, and 3.15, respectively. Plots for the corresponding approximations can be found in Figures 3.14, 3.15, and 3.16. Note, while we observe the theoretical first order bound for the approximation error, we also observe that the higher order elements yield more accurate approximations.

Table 3.10: Rates of convergence for Example 3.7 using $r = 0$, $\alpha = 2I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h	L^∞ norm	order	L^2 norm	order
5.55e-01	2.59e-01		2.73e-01	
3.70e-01	1.63e-01	1.14	1.75e-01	1.10
2.78e-01	1.17e-01	1.17	1.29e-01	1.06
1.85e-01	7.29e-02	1.16	8.48e-02	1.03
1.39e-01	5.27e-02	1.13	6.33e-02	1.02

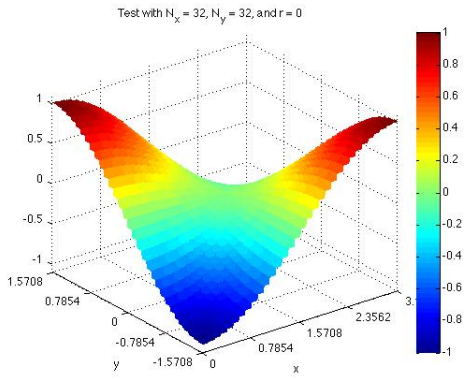


Figure 3.11: Computed solution for Example 3.7 using $r = 0$, $\alpha = 2I$, $h = 1.388\text{e-}01$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

Table 3.11: Rates of convergence for Example 3.7 using $r = 1$, $\alpha = 2I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h	L^∞ norm	order	L^2 norm	order
5.55e-01	4.89e-02		2.84e-02	
3.70e-01	2.23e-02	1.93	1.29e-02	1.94
2.78e-01	1.27e-02	1.97	7.38e-03	1.95

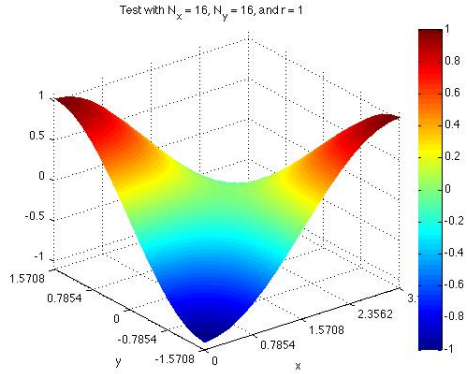


Figure 3.12: Computed solution for Example 3.7 using $r = 1$, $\alpha = 2I$, $h =$ 2.777e-01, and *fsolve* with initial guess $u_h^{(0)} = 0$.

Table 3.12: Rates of convergence for Example 3.7 using $r = 2$, $\alpha = 2I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h	L^∞ norm	order	L^2 norm	order
2.22e+00	2.82e-01		1.25e-01	
7.40e-01	9.04e-03	3.13	9.52e-03	2.35
4.44e-01	2.39e-03	2.60	2.88e-03	2.34

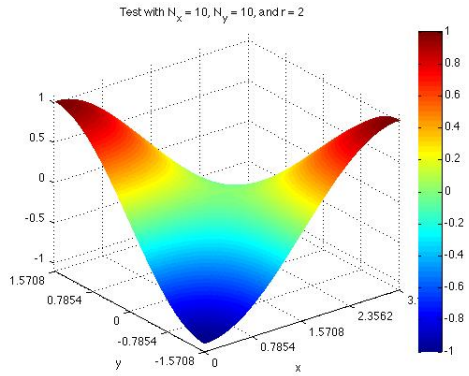


Figure 3.13: Computed solution for Example 3.7 using $r = 2$, $\alpha = 2I$, $h =$ 4.443e-01, and *fsolve* with initial guess $u_h^{(0)} = 0$.

Table 3.13: Rates of convergence for Example 3.8 using $r = 0$, $\alpha = 60I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h	L^∞ norm	order	L^2 norm	order
2.83e-01	4.50e-01		3.37e-01	
1.41e-01	2.83e-01	0.67	2.02e-01	0.74
1.18e-01	2.46e-01	0.78	1.72e-01	0.88
9.43e-02	2.05e-01	0.82	1.40e-01	0.93

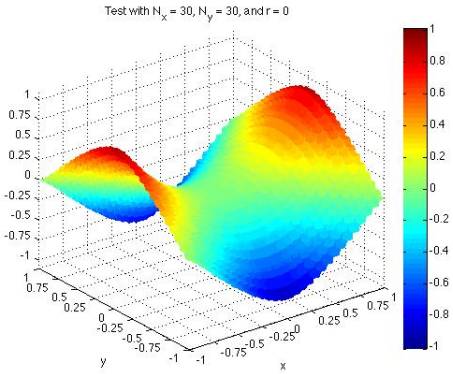


Figure 3.14: Computed solution for Example 3.8 using $r = 0$, $\alpha = 60I$, $h = 9.428\text{e-}02$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

Table 3.14: Rates of convergence for Example 3.8 using $r = 1$, $\alpha = 60I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h	L^∞ norm	order	L^2 norm	order
4.71e-01	4.36e-02		3.17e-02	
2.83e-01	2.79e-02	0.88	1.81e-02	1.09
2.02e-01	2.20e-02	0.71	1.29e-02	1.02

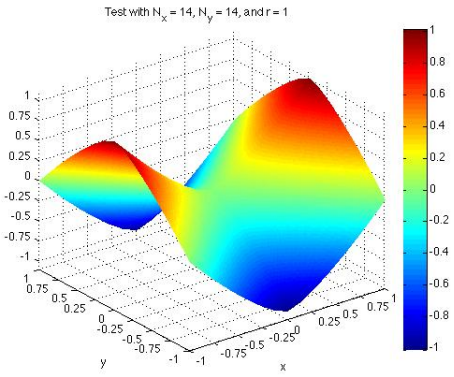


Figure 3.15: Computed solution for Example 3.8 using $r = 1$, $\alpha = 60I$, $h =$ 2.020e-01, and *fsolve* with initial guess $u_h^{(0)} = 0$.

Table 3.15: Rates of convergence for Example 3.8 using $r = 2$, $\alpha = 60I$, and *fsolve* with initial guess $u_h^{(0)} = 0$.

h	L^∞ norm	order	L^2 norm	order
5.66e-01	2.41e-02		8.71e-03	
4.71e-01	1.48e-02	2.66	7.58e-03	0.76
3.54e-01	1.06e-02	1.16	4.64e-03	1.71

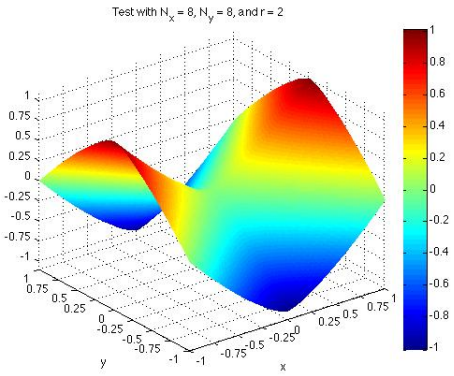


Figure 3.16: Computed solution for Example 3.8 using $r = 2$, $\alpha = 60I$, $h =$ 3.536e-01, and *fsolve* with initial guess $u_h^{(0)} = 0$.

3.5.3 The Role of the Numerical Moment

In this section, we focus on understanding the role of the numerical moment through numerical experiments. Using a standard discretization technique from numerical PDE theory for linear equations can lead to the presence of numerical artifacts, i.e., algebraic solutions of the system of equations that do not correspond to a PDE solution (see Section 1.3 for more details). We show that using a numerical moment can eliminate these numerical artifacts in many instances, and in certain instances when the numerical artifacts are not fully eliminated, the algebraic system has enough structure to design solvers that are consistent in searching for viscosity solutions of the nonlinear PDE problem. Thus, the presence of numerical artifacts can be handled at the solver level using the numerical moment when such algebraic solutions do exist.

Our main emphasis will be on the Monge-Ampère type problem from Example 3.1. The given problem has two classical PDE solutions, u^+ and u^-. However, there exists infinitely many C^1 functions that satisfy the PDE and boundary conditions almost everywhere with respect to the Lebesgue measure, as seen by $\widehat{\mu}$ defined below in (3.44). These almost everywhere solutions will correspond to what we call numerical artifacts in that algebraic solutions for a given discretization may correspond to these functions. It is well known that using a standard linear discretization scheme for the Monge-Ampère problem can yield multiple solutions, many of which are numerical artifacts that do not correspond to PDE solutions (cf. [27]). For example, let $\widehat{\mu} \in H^2(0,1) \setminus C^2(0,1)$ be defined by

$$\widehat{\mu}(x) = \begin{cases} \frac{1}{2}x^2 + \frac{1}{4}x, & \text{for } x < 0.5, \\ -\frac{1}{2}x^2 + \frac{5}{4}x - \frac{1}{4}, & \text{for } x \geq 0.5. \end{cases} \tag{3.44}$$

Furthermore, suppose $x = 0.5$ is a node in the partition. Then, when using a standard LDG or IPDG discretization, $\widehat{\mu}$ corresponds to a numerical solution, yielding a numerical artifact.

We now compare our nonstandard LDG discretizations that use a numerical moment to LDG methods without a numerical moment. Observe, when $\alpha = 0$, we have no numerical moment. As a result, we have numerical artifacts as in the standard LDG or IPDG discretization case. Suppose $r = 0$. Then, for $\alpha \neq 0$, inspection of (3.20) yields the fact that $\frac{P_h^{+-}+P_h^{-+}}{2}$ cannot jump from a value of 1 to a value of -1 when crossing $x = 0.5$. Thus, the numerical moment penalizes discontinuities in $P_h^{\mu\nu}$, $\mu, \nu \in \{+, -\}$, and, as a result, the numerical moment eliminates numerical artifacts such as $\widehat{\mu}$. Similarly, for $r = 1$, we can see that $\widehat{\mu}$ does not correspond to a numerical solution. However, in this case, the algebraic system does have a small residual well that may trap solvers such as *fsolve*. Thus, for $r = 0$ and $r = 1$, the numerical moment penalizes differences in $P_h^{\mu\nu}$, $\mu, \nu \in \{+, -\}$, that arise from discontinuities in u_h, q_h^+, and q_h^-. Hence, it eliminates numerical artifacts such as $\widehat{\mu}$.

Next, we consider $r \geq 2$, in which case $\widehat{\mu} \in V^h$. Furthermore, since $\widehat{\mu} \in C^1(\Omega)$, we will end up with $u_h = \widehat{\mu}$, $q_h^+ = q_h^-$, and $P_h^{\mu\nu}$, $\mu, \nu \in \{+, -\}$, being a numeric solution, where

$$q_h^+(x) = \begin{cases} x + \frac{1}{4}, & \text{for } x < 0.5, \\ -x + \frac{5}{4}, & \text{for } x > 0.5, \end{cases} \qquad \text{and} \qquad P_h^{+-}(x) = \begin{cases} 1, & \text{for } x < 0.5, \\ -1, & \text{for } x > 0.5. \end{cases}$$

Thus, by the equalities of $P_h^{\mu\nu}$, $\mu, \nu \in \{+, -\}$, the numerical moment is always zero and we do have numerical artifacts. These equalities are a consequence of the continuity of u_h, q_h^+, and q_h^-. With the extra degrees of freedom for $r \geq 2$, we allow C^1 to be embedded into our approximation space V^h for nontrivial functions, thus creating possible solutions with a zero valued numerical moment. The numerical moment acts as a penalty term for differences in $P_h^{\mu\nu}$, $\mu, \nu \in \{+, -\}$, which are a consequence of discontinuities in q_h^+ and q_h^- that naturally arise for nontrivial functions when $r = 0$ or $r = 1$.

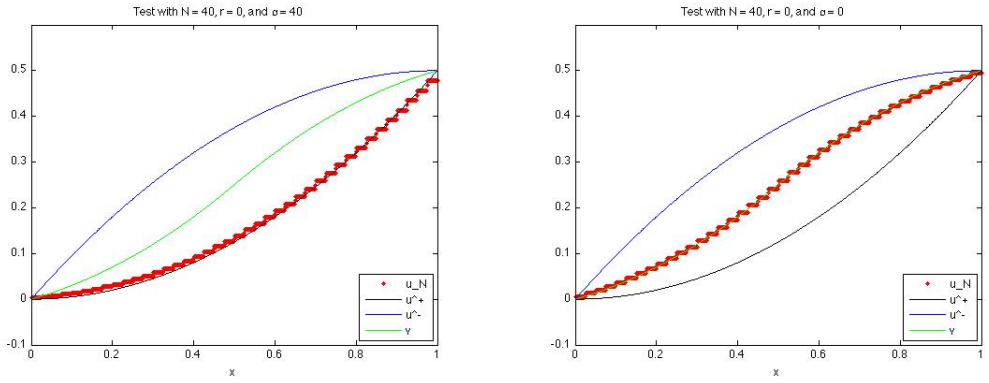
Even with the possible presence of numerical artifacts for the above discretization when $r \geq 2$, the numerical moment can be exploited at the solver level, such as in

Algorithm 3.1. For the next numerical tests, we will show that using Algorithm 3.1 with a sufficiently large coefficient for the numerical moment destabilizes numerical artifacts such as $\widehat{\mu}$ and steers the approximation towards the viscosity solution of the PDE. Let $\overline{u}(x) = \frac{x}{2}$. Then, $\overline{u}$ is the secant line formed by the boundary data for the given boundary value problem. We now approximate the solution of the Monge-Ampère type problem from Example 3.1 by using 100 iterations of Algorithm 3.1 followed by using *fsolve* on the mixed formulation to solve the global discretization given by (3.12) and (3.16). We take the initial guess to be

$$u_h^{(0)} = \frac{3}{4}\widehat{\mu} + \frac{1}{4}\overline{u},$$

where, for $r = 0$, u_h^0 is first projected into V^h. From Figure 3.17, we see that the numerical moment drives the solution towards the viscosity solution u^+ when $r = 0$ and α is positive. From Figure 3.18, we see that the numerical moment also drives the solution towards the viscosity solution u^+ when $r = 2$ and α is positive, despite the presence of numerical artifacts. From Figure 3.19, we see that the numerical moment drives the solution towards the viscosity solution of $F(u_{xx}, u_x, u, x) := u_{xx}^2 - 1 = 0$ given by u^- for $r = 0$ and $r = 2$ when α is chosen to be negative. In each figure, the middle graph corresponds to $\widehat{\mu}$. Clearly, we recover the numerical artifact corresponding to $\widehat{\mu}$ when $\alpha = 0$. Thus, the numerical moment plays an essential role in either eliminating numerical artifacts at the discretization level or handling numerical artifacts at the solver level.

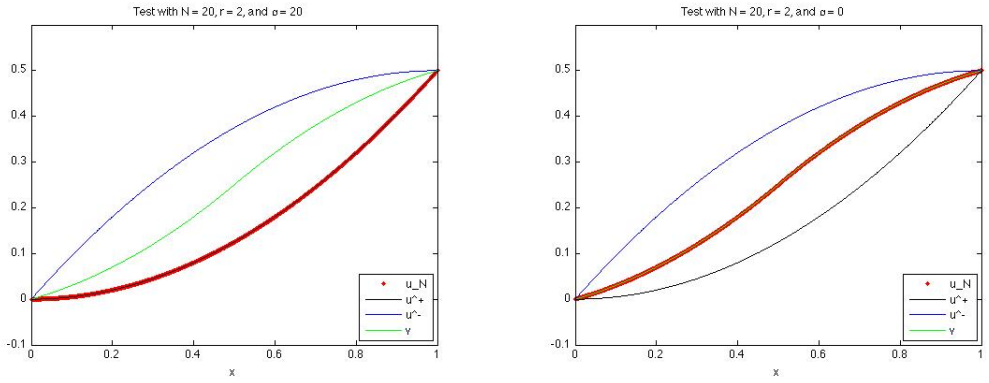
As another example of the numerical moment assisting with the issue of numerical artifacts and uniqueness only in a restrictive function class, we approximate Example 3.5 using the numerical moment with $\alpha = -12\mathbf{1}$, $N_x = N_y = 24$, $r = 0$, and initial guess given by the zero function. The result is recorded in Figure 3.20. Thus, we can see that for a negative semi-definite choice for α, we recover an approximation for the non-convex solution of the Monge-Ampère problem represented in Figure 3.5.



(a) $\alpha = \frac{1}{h}$. **(b)** $\alpha = 0$.

Figure 3.17: Computed solutions for Example 3.1 using $r = 0$, $h =$ 2.500e-02, and Algorithm 3.1 with initial guess $u_h^{(0)} = \mathcal{P}_h\left(\frac{3}{4}\widehat{\mu} + \frac{1}{4}\overline{u}\right)$.

Another benefit of the numerical moment is that it can help regularize a problem that may not be well-conditioned for a Newton solver due to a singular or poorly scaled Jacobian. We again consider the problem given in Example 3.6. Note that $\frac{\partial F}{\partial D^2 u} = 0$ almost everywhere in Ω for the viscosity solution u because $D^2u(x, y) = 0$ for all $x \neq 0$. This leads to a singular or badly scaled matrix when using a Newton algorithm to solve the problem without the presence of a numerical moment. By adding a numerical moment, the resulting system of equations may be better suited for Newton algorithms since $\frac{\partial \widehat{F}}{\partial P_h^{\pm\mp}} = \frac{\partial F}{\partial P_h^{\pm\mp}} - \alpha$ may be nonsingular even when $P_h^{\pm\mp} \approx 0$. For the next numerical test, we let $\alpha = \gamma\mathbf{1}$ for various positive values of γ to see how the numerical moment affects both the accuracy and the performance of the Newton solver *fsolve*. The choice for the numerical moment is especially interesting upon noting that α is in fact a singular matrix. However, with a numerical moment, the perturbation in $\frac{\partial \widehat{F}}{\partial P_h^{\pm\mp}}$ caused by $P_h^{\pm\mp}$ may be enough to eliminate the singularity due to the fact the approximation may now have some curvature. We let the initial guess be given by the zero function, fix the mesh $N_x = N_y = 20$, and let $r = 0$. We can see from Table 3.16 that for γ small, *fsolve* converges slowly, if at all. For $\gamma = 0$, *fsolve* does not converge within 100 iterations even for a very good initial guess. However,



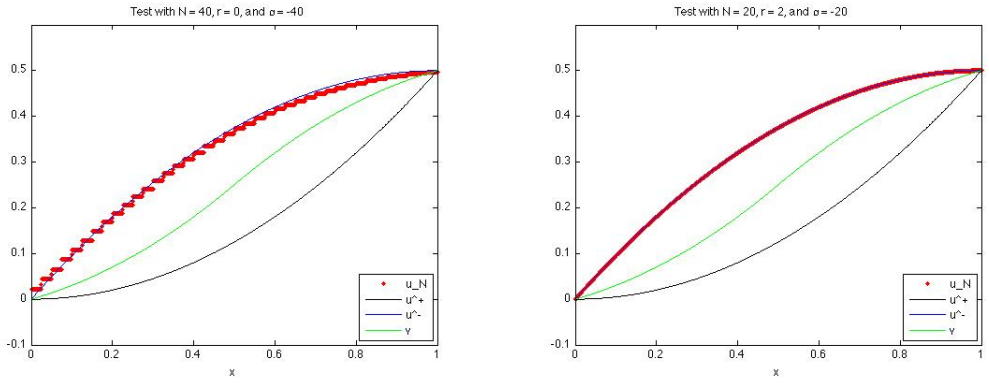
(a) $\alpha = \frac{1}{h}$. **(b)** $\alpha = 0$.

Figure 3.18: Computed solutions for Example 3.1 using $r = 2$, $h =$ 5.000e-02, and Algorithm 3.1 with initial guess $u_h^{(0)} = \frac{3}{4}\widehat{\mu} + \frac{1}{4}\overline{u}$.

increasing γ does appear to aid *fsolve* in its ability to find a root with only a small penalty in the approximation error. For $r \geq 1$, we again note that Algorithm 3.1 provides a much better suited solver due to the degeneracy of the problem. However, the crux of Algorithm 3.1 reduces to a choice of $\gamma > 0$ with $\alpha = \gamma I$ instead of $\alpha = \gamma \mathbf{1}$. Similar results, as seen in Table 3.16, hold for $\alpha = \gamma I$.

Table 3.16: Approximation errors when varying $\alpha = \gamma \mathbf{1}$ for Example 3.6 using $r = 0$, $h =$ 7.071e-02, and *fsolve* with initial guess $u_h^{(0)} = 0$. The entry 0^* corresponds to an initial guess given by the L^2-projection of $u(x, y) = |x|$, $q_h^{\pm}(x, y) = \text{sgn}(x)$, and $P_h^{\mu\nu}(x, y) = 0$ for $\mu, \nu \in \{+, -\}$. The nonlinear solver *fsolve* is set to perform a maximum of 100 iterations.

γ	L^∞ norm	L^2 norm	*fsolve* iterations
600	2.43e-01	2.43e-01	9
60	2.29e-01	2.27e-01	9
12	2.02e-01	1.98e-01	10
4	1.81e-01	1.74e-01	10
1	3.40e-01	2.08e-01	100
0^*	2.84e-01	1.96e-01	100



(a) $r = 0$ and $h = 2.500\text{e-}02$. **(b)** $r = 2$ and $h = 5.000\text{e-}02$.

Figure 3.19: Computed solutions for Example 3.1 using $\alpha = -\frac{1}{h}$ and Algorithm 3.1 with initial guess $u_h^{(0)} = \mathcal{P}_h\left(\frac{3}{4}\widehat{\mu} + \frac{1}{4}\overline{u}\right)$.

We make one final note about using the iterative solver given by Algorithm 3.1. Note that using *fsolve* to solve the full system with the initial guess given by $u_h^{(0)} = 0$ resulted in either not finding a root for many tests ($r = 1$) or converging to a numerical artifact with a discontinuous second order derivative at another node in the mesh ($r = 2$). In order to use *fsolve* for the given test problem, the initial guess should either be restricted to the class of functions where P_h^{+-} and P_h^{-+} preserve the ellipticity of the nonlinear operator, the initial guess should be preconditioned by first using *fsolve* with $r = 0, 1$, or the initial guess should be preconditioned using Algorithm 3.1. When using $r = 0$ and a non-ellipticity-preserving initial guess, solving the full system of equations with *fsolve* still has the potential to converge to u^- in Example 3.1 even for $\alpha > 0$. The strength of Algorithm 3.1 is that it strongly enforces the requirement that $\widehat{F}$ is monotone decreasing in P_h^{+-} and P_h^{-+} over each iteration. Thus, a sufficiently large value for α drives the approximation towards the class of ellipticity-preserving functions if the algorithm converges.

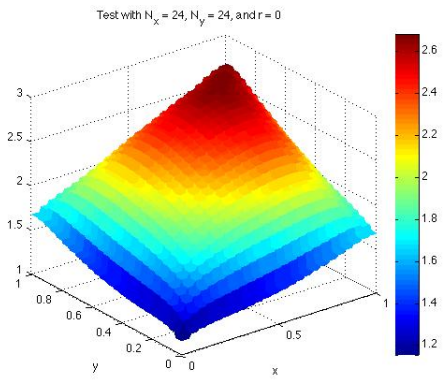


Figure 3.20: Computed solution for Example 3.5 using $r = 0$, $\alpha = -121$, $h =$ 5.893e-02, and *fsolve* with initial guess $u_h^{(0)} = 0$.

3.5.4 Parabolic Problems

We now implement some of the proposed fully discrete DG methods for approximating the fully nonlinear parabolic problem (3.2). We will use RK4 and trapezoidal DG methods for approximating problems in one spatial dimension and backwards Euler DG methods for approximating problems in two spatial dimensions. While the above formulation makes no attempt to formally quantify a CFL condition for the RK4 method, for the tests we assume a CFL constraint of the form $\Delta t = \kappa_t h^2$ and note that the constant κ_t appears to decrease as the order of the elements increases. However, we make no attempt to classify and compare the efficiency of the various time-discretization methods. Instead, we focus on testing and demonstrating the usability of the fully discrete schemes and their promising potentials. For explicit scheme tests we record the parameter κ_t, and for implicit scheme tests we record the time step size Δt. Note that the row 0^* in the tables corresponding to the RK4 method refers to elements with $r = 0$ that use the standard L^2-projection operator in (3.38), i.e., $\delta = 0$. Otherwise, we set $\delta = \frac{1}{h}$.

Example 3.9. *Consider the problem*

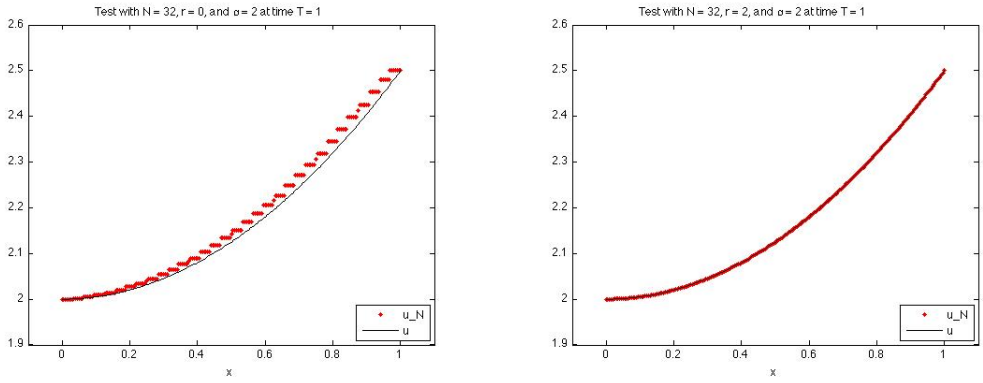
$$\begin{aligned} u_t - u_{xx}\, u &= f && \text{in } \Omega \times (0,1], \\ u &= g && \text{on } \partial\Omega \times (0,1], \\ u &= u_0 && \text{in } \Omega \times \{0\}, \end{aligned}$$

where $\Omega = (0,1)$, $f(x,t) = -\frac{1}{2}x^2 - t^4 + 4\,t^3 - 1$, *and* g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = 0.5\,x^2 + t^4 + 1$.

Notice that the PDE is actually quasi-linear, but it does provide a measure of the effectiveness of the implementation. The numerical results for RK4 are presented in Table 3.17 and Figure 3.21, and the results for the trapezoidal method are shown in Table 3.18 and Figure 3.22. As expected, both schemes have high levels of accuracy when using $r = 2$. In fact, RK4 has potential to recover the exact solution when using $r = 2$ due to the fourth order time-stepping. Thus, we are able to gauge the potential for our implementation.

Table 3.17: Rates of convergence in space for Example 3.9 at time $t = 1$ using RK4 time-stepping with $\alpha = 2$, $\kappa_t = 0.001$, and $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = 1/4$	$h = 1/8$		$h = 1/16$		$h = 1/32$	
		Error	Error	Order	Error	Order	Error	Order
0	L^2	9.9e-02	6.4e-02	0.64	3.6e-02	0.81	1.9e-02	0.92
	L^∞	2.2e-01	1.4e-01	0.64	8.0e-02	0.81	4.3e-02	0.89
1	L^2	5.7e-03	1.5e-03	1.98	3.7e-04	1.99	9.2e-05	1.99
	L^∞	8.0e-03	2.0e-03	1.99	5.1e-04	1.99	1.3e-04	1.99
2	L^2	2.4e-08	2.4e-08	0.00	2.4e-08	0.00	2.4e-08	-0.00
	L^∞	3.6e-08	3.7e-08	-0.03	3.7e-08	-0.01	3.7e-08	-0.01



(a) $r = 0$. **(b)** $r = 2$.

Figure 3.21: Computed solutions at time $t = 1$ for Example 3.9 using RK4 time-stepping with $\alpha = 2$, $h =$ 3.125e-02, $\kappa_t = 0.001$, and $u_h^0 = \mathcal{P}_h u_0$.

Example 3.10. *Consider the problem*

$$\begin{aligned} u_t - u_x \ln\big(u_{xx} + 1\big) &= f \qquad \text{in } \Omega \times (0, 1/2], \\ u &= g \qquad \text{on } \partial\Omega \times (0, 1/2], \\ u &= u_0 \qquad \text{in } \Omega \times \{0\}, \end{aligned}$$

where $\Omega = (0, 2)$, $f(x,t) = -e^{(t+1)x}\Big(x - (t+1)\ln\big((t+1)^2 e^{(t+1)x} + 1\big)\Big)$, *and* g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = e^{(t+1)x}$.

Notice, the exact solution u cannot be factored into the form $u(x,t) = G(t)\,Y(x)$ for some functions G and Y. Results for RK4 are given in Table 3.19 and Figure 3.23, and results for the trapezoidal method are shown in Table 3.20 and Figure 3.24. We note that RK4 was unstable without using the very restrictive values for κ_t recorded in Table 3.19. However, for RK4, we observe optimal rates of convergence in the spatial variable while the rates for the trapezoidal method appear to be limited by the lower rate of convergence for the time-stepping scheme.

Table 3.18: Rates of convergence in space for Example 3.9 at time $t = 1$ using trapezoidal time-stepping with $\alpha = 2$, $\Delta t = 0.001$, and *fsolve* with initial guess $u_h^{(0)} = \mathcal{P}_h u_0$.

r	Norm	$h = 1/4$	$h = 1/8$		$h = 1/16$		$h = 1/32$	
		Error	Error	Order	Error	Order	Error	Order
0	L^2	5.9e-02	3.4e-02	0.78	1.9e-02	0.84	1.0e-02	0.91
	L^∞	1.9e-01	1.1e-01	0.79	5.9e-02	0.90	3.0e-02	0.95
1	L^2	5.0e-03	1.4e-03	1.86	3.6e-04	1.94	9.1e-05	1.97
	L^∞	1.1e-02	2.8e-03	2.00	7.1e-04	2.00	1.8e-04	2.00
2	L^2	1.1e-07	1.1e-07	0.00	1.1e-07	0.00	1.1e-07	0.00
	L^∞	1.5e-07	1.6e-07	-0.04	1.6e-07	-0.01	1.6e-07	-0.00

Example 3.11. *Consider the problem*

$$\begin{aligned} u_t - \min_{\theta(t,x)\in\{1,2\}} \Big\{ A_\theta \, u_{xx} - c(x,t)\cos(t)\,\sin(x) - \sin(t)\,\sin(x) \Big\} = 0 \quad & \text{in } \Omega \times (0, \frac{1}{2}], \\ u = g \quad & \text{on } \partial\Omega \times (0,1], \\ u = u_0 \quad & \text{in } \Omega \times \{0\}, \end{aligned}$$

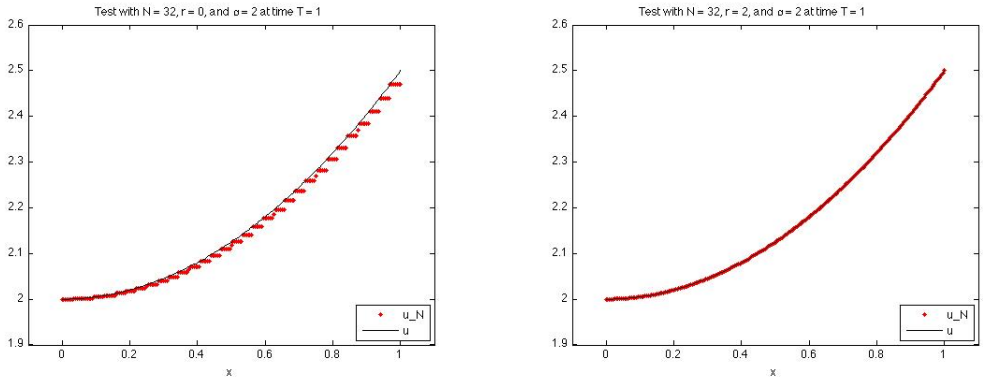
where $\Omega = (0, 2\pi)$, $A_1 = 1, A_2 = \frac{1}{2}$,

$$c(x,t) = \begin{cases} 1, & \text{if } 0 < t \le \frac{\pi}{2} \text{ and } 0 < x \le \pi \text{ or } \frac{\pi}{2} < t \le \pi \text{ and } \pi < x < 2\pi, \\ \frac{1}{2}, & \text{otherwise,} \end{cases}$$

and g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = \cos(t)\sin(x)$.

Notice that this problem involves an optimization over a finite dimensional set, and the solution corresponds to

$$\theta(x,t) = \begin{cases} 1, & \text{if } c(x,t) = 1, \\ 2, & \text{if } c(x,t) = \frac{1}{2}. \end{cases}$$



(a) $r = 0$. **(b)** $r = 2$.

Figure 3.22: Computed solutions at time $t = 1$ for Example 3.9 using trapezoidal time-stepping with $\alpha = 2$, $h =$ 3.125e-02, $\Delta t = 0.001$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

The numerical results are reported in Table 3.21 and Figure 3.25 for RK4 and in Table 3.22 and Figure 3.26 for the trapezoidal method.

Example 3.12. *Consider the problem*

$$\begin{aligned} u_t - \inf_{-1\le\theta(t,x)\le 1}\Big\{|x-1|u_{xx} + \theta u_x\Big\} &= f \qquad \text{in } \Omega\times(0,\frac{1}{2}],\\ u &= g \qquad \text{on } \partial\Omega\times(0,1],\\ u &= u_0 \qquad \text{in } \Omega\times\{0\}, \end{aligned}$$

where $\Omega = (0,3)$, $f(x,t) = -|x-1|^2\,(|x-1|+3)\,e^{-t}$, *and* g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = |x-1|^3\,e^{-t}$.

Notice that the above operator is not second order when $x = 1$. For each value of t, we have $u \in H^3(0,3)$. Also, the solution corresponds to

$$\theta(x,t) = \begin{cases} 1, & \text{if } x < 1,\\ -1, & \text{if } x > 1. \end{cases}$$

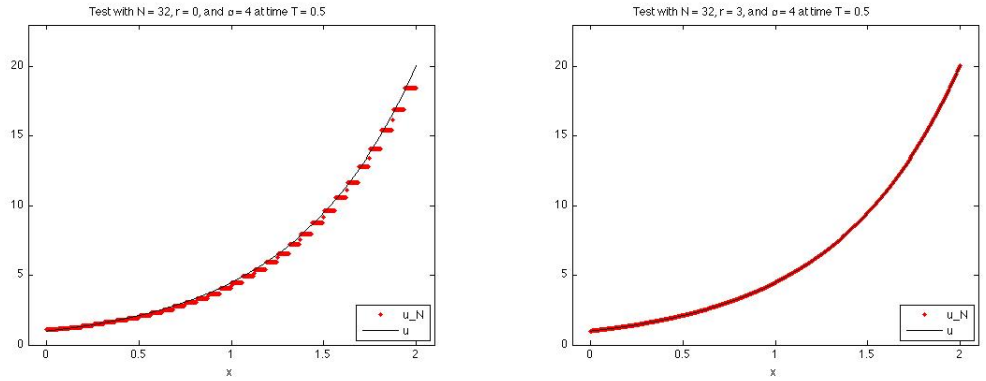
Table 3.19: Rates of convergence in space for Example 3.10 at time $t = 0.5$ using RK4 time-stepping with $\alpha = 4$, $\kappa_t = 0.005, 0.001, 0.0005, 0.0001$ for $r = 0, 1, 2, 3$, respectively, and $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = 2/4$	$h = 2/8$		$h = 2/16$		$h = 2/32$	
		Error	Error	Order	Error	Order	Error	Order
0	L^2	6.9e+00	5.7e+00	0.28	4.1e+00	0.47	2.6e+00	0.65
	L^∞	1.0e+01	7.9e+00	0.40	5.6e+00	0.50	3.7e+00	0.60
0^*	L^2	2.8e+00	1.5e+00	0.89	8.6e-01	0.82	5.1e-01	0.76
	L^∞	7.8e+00	5.0e+00	0.65	2.9e+00	0.76	1.6e+00	0.85
1	L^2	4.8e-01	1.2e-01	2.05	3.0e-02	1.93	8.2e-03	1.88
	L^∞	8.4e-01	2.3e-01	1.87	5.8e-02	1.99	1.5e-02	1.92
2	L^2	3.7e-02	7.8e-03	2.23	1.9e-03	2.03	4.8e-04	2.00
	L^∞	5.9e-02	1.0e-02	2.53	2.2e-03	2.19	5.2e-04	2.11
3	L^2	1.1e-03	7.4e-05	3.95	4.7e-06	3.99	3.01e-07	3.96
	L^∞	2.5e-03	1.6e-04	3.93	1.1e-05	3.86	7.66e-07	3.84

We can see from the results for the RK4 method in Table 3.23 and Figure 3.27 and the results for the trapezoidal method in Table 3.24 and Figure 3.28 that the spatial rates of convergence appear to be limited by two instead of the optimal rate of three for $r \geq 2$, while the accuracy appears to increase with respect to the degree of the elements.

Example 3.13. *Consider the dynamic Hamilton-Jacobi-Bellman problem*

$$\begin{aligned} u_t + \min\{-\Delta u, -\Delta u/2\} &= f && \text{in } \Omega \times (0,1], \\ u &= g && \text{on } \partial\Omega \times (0,1], \\ u &= u_0 && \text{in } \Omega \times \{0\}, \end{aligned}$$



(a) $r = 0$ with $\kappa_t = 0.005$ and $\delta = 0$. **(b)** $r = 3$ with $\kappa_t = 0.0001$ and $\delta = \frac{1}{h}$.

Figure 3.23: Computed solutions at time $t = 0.5$ for Example 3.10 using RK4 time-stepping with $\alpha = 4$, $h =$ 6.250e-02, and $u_h^0 = \mathcal{P}_h u_0$.

where $\Omega = (-1,1) \times (-1,1)$, $f(x,y,t) = s(x,y,t) + 2\,t\,(x\,|x| + y\,|y|)$,

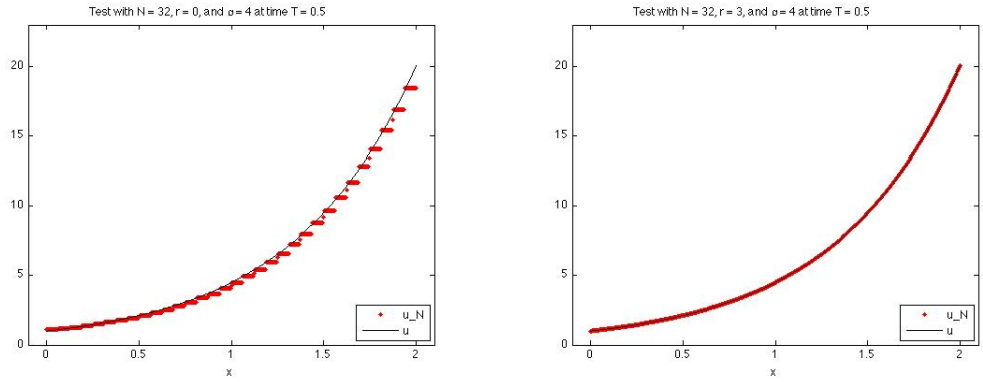
$$s(x,y,t) = \begin{cases} 2t^2, & \text{if } x < 0 \text{ and } y < 0, \\ -4t^2, & \text{if } x > 0 \text{ and } y > 0, \\ 0, & \text{otherwise,} \end{cases}$$

and g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,y,t) = t^2\,x\,|x| + t\,y\,|y|$. *Then, for all* t, *we have* $u(\cdot,\cdot,t) \in H^2(\Omega)$.

We expect the spatial rate of convergence to be bounded by 2. However, due to the low order time discretization scheme, we can see that our error is dominated by the time discretization for $r \geq 1$. The spatial orders of convergence for $r = 0$ and $r = 1$ are recorded in Tables 3.25 and 3.26, respectively. For $r = 0$, the spatial discretization order matches the time discretization order, and we do observe an optimal rate of convergence. Using $r = 2$, we have the solution $u \in V^h$. Due to the high level of accuracy when using $r = 2$, we observe that the time discretization order is in fact 1 as shown in Table 3.27. Plots for some of the approximations can be found in Figures 3.29, 3.30, and 3.31.

Table 3.20: Rates of convergence in space for Example 3.10 at time $t = 0.5$ using trapezoidal time-stepping with $\alpha = 4$, $\Delta t = 0.005$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = 2/4$	$h = 2/8$		$h = 2/16$		$h = 2/32$	
		Error	Error	Order	Error	Order	Error	Order
0	L^2	2.8e+00	1.5e+00	0.89	8.6e-01	0.82	5.1e-01	0.77
	L^∞	7.8e+00	5.0e+00	0.65	2.9e+00	0.76	1.6e+00	0.85
1	L^2	3.8e-01	1.3e-01	1.62	4.5e-02	1.49	1.5e-02	1.60
	L^∞	1.3e+00	4.0e-01	1.74	1.1e-01	1.85	3.0e-02	1.91
2	L^2	2.7e-02	6.7e-03	2.04	1.9e-03	1.82	5.3e-04	1.85
	L^∞	1.0e-01	1.5e-02	2.77	2.1e-03	2.80	5.4e-04	1.99
3	L^2	1.1e-03	7.2e-05	3.89	1.3e-05	2.46	1.2e-05	0.09
	L^∞	5.5e-03	4.0e-04	3.76	2.7e-05	3.92	1.3e-05	1.05

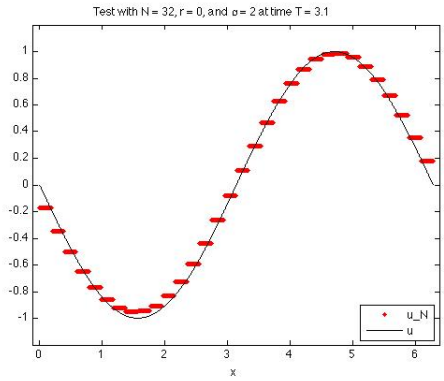


(a) $r = 0$. **(b)** $r = 3$.

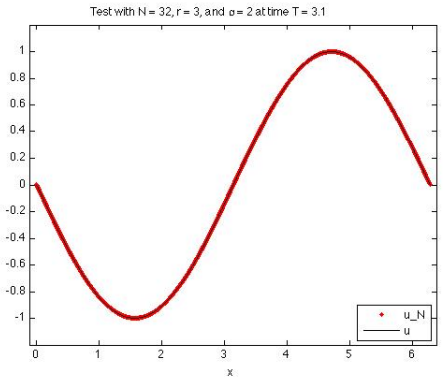
Figure 3.24: Computed solutions at time $t = 0.5$ for Example 3.10 using trapezoidal time-stepping with $\alpha = 4$, $h = 6.250\text{e-}02$, $\Delta t = 0.005$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

Table 3.21: Rates of convergence in space for Example 3.11 at time $t = 3.1$ using RK4 time-stepping with $\alpha = 2$, $\kappa_t = 0.05, 0.005, 0.001, 0.0005$ for $r = 0, 1, 2, 3$, respectively, and $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = \pi/2$	$h = \pi/4$		$h = \pi/8$		$h = \pi/16$	
		Error	Error	Order	Error	Order	Error	Order
0	L^2	1.5e+00	1.3e+00	0.24	7.1e-01	0.82	3.3e-01	1.12
	L^∞	9.6e-01	7.7e-01	0.31	4.9e-01	0.65	2.9e-01	0.78
0*	L^2	1.2e+00	6.9e-01	0.78	3.0e-01	1.19	1.4e-01	1.16
	L^∞	9.2e-01	5.8e-01	0.67	3.2e-01	0.85	1.8e-01	0.83
1	L^2	2.7e-01	7.6e-02	1.81	2.0e-02	1.94	5.0e-03	1.99
	L^∞	1.9e-01	6.6e-02	1.52	1.7e-02	1.97	4.2e-03	2.00
2	L^2	7.2e-02	1.8e-02	1.98	4.5e-03	2.02	1.1e-03	2.01
	L^∞	6.8e-02	1.8e-02	1.93	4.3e-03	2.04	1.1e-03	2.03
3	L^2	8.3e-03	5.7e-04	3.87	3.6e-05	3.97	2.2e-06	4.02
	L^∞	7.6e-03	5.1e-04	3.92	3.2e-05	4.00	1.9e-06	4.04



(a) $r = 0$ with $\kappa_t = 0.05$ and $\delta = 0$.

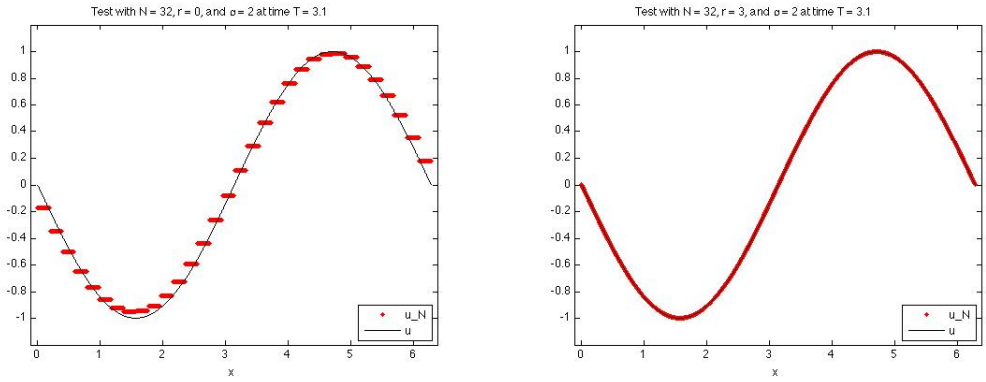


(b) $r = 3$ with $\kappa_t = 0.0005$ and $\delta = \frac{1}{h}$.

Figure 3.25: Computed solutions at time $t = 3.1$ for Example 3.11 using RK4 time-stepping with $\alpha = 2$, $h = 1.963\text{e-}01$, and $u_h^0 = \mathcal{P}_h u_0$.

Table 3.22: Rates of convergence in space for Example 3.11 at time $t = 3.1$ using trapezoidal time-stepping with $\alpha = 2$, $\Delta t = 0.031$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = \pi/2$	$h = \pi/4$		$h = \pi/8$		$h = \pi/16$	
		Error	Error	Order	Error	Order	Error	Order
0	L^2	1.2e+00	6.9e-01	0.78	3.0e-01	1.19	1.4e-01	1.16
	L^∞	9.2e-01	5.8e-01	0.67	3.2e-01	0.85	1.8e-01	0.83
1	L^2	2.3e-01	7.5e-02	1.63	2.0e-02	1.89	7.9e-03	1.36
	L^∞	2.1e-01	6.6e-02	1.66	1.7e-02	1.95	5.8e-03	1.56
2	L^2	6.9e-02	1.8e-02	1.93	4.7e-03	1.95	2.5e-03	0.90
	L^∞	6.5e-02	1.8e-02	1.87	4.5e-03	2.00	1.9e-03	1.21
3	L^2	8.1e-03	5.9e-04	3.79	1.1e-04	2.42	1.1e-04	0.03
	L^∞	7.5e-03	5.2e-04	3.87	7.6e-05	2.77	7.3e-05	0.06

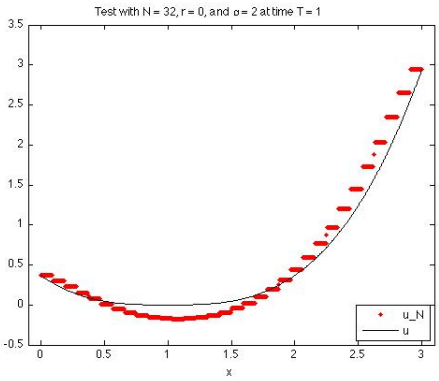


(a) $r = 0$. **(b)** $r = 3$.

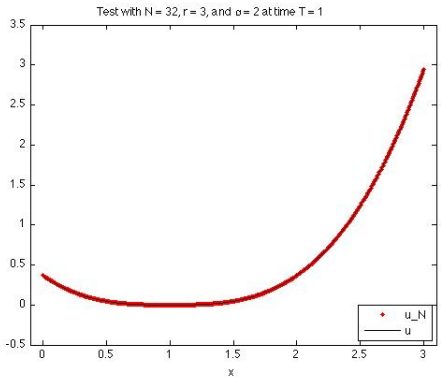
Figure 3.26: Computed solutions at time $t = 3.1$ for Example 3.11 using trapezoidal time-stepping with $\alpha = 2$, $h = 1.963\text{e-}01$, $\Delta t = 0.031$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

Table 3.23: Rates of convergence in space for Example 3.12 at time $t = 1$ using RK4 time-stepping with $\alpha = 2$, $\kappa_t = 0.05, 0.005, 0.001, 0.0005$ for $r = 0, 1, 2, 3$, respectively, and $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = 3/4$	$h = 3/8$		$h = 3/16$		$h = 3/32$	
		Error	Error	Order	Error	Order	Error	Order
0	L^2	2.1e+00	1.4e+00	0.51	5.3e-01	1.44	2.9e-01	0.88
	L^∞	2.1e+00	1.5e+00	0.44	8.5e-01	0.86	4.9e-01	0.81
0*	L^2	8.6e-01	8.8e-01	-0.02	5.9e-01	0.58	2.9e-01	1.00
	L^∞	1.8e+00	1.3e+00	0.53	7.3e-01	0.80	3.9e-01	0.92
1	L^2	2.0e-01	7.3e-02	1.45	1.3e-02	2.44	3.2e-03	2.06
	L^∞	3.9e-01	1.0e-01	1.92	2.7e-02	1.95	6.8e-03	1.98
2	L^2	1.1e-01	2.0e-02	2.49	5.2e-03	1.98	1.3e-03	2.02
	L^∞	1.1e-01	2.0e-02	2.42	5.2e-03	1.99	1.3e-03	2.00
3	L^2	3.0e-02	6.8e-03	2.13	1.7e-03	2.00	4.2e-04	2.03
	L^∞	2.9e-02	6.9e-03	2.08	1.7e-03	2.02	4.2e-04	2.02



(a) $r = 0$ with $\kappa_t = 0.05$.

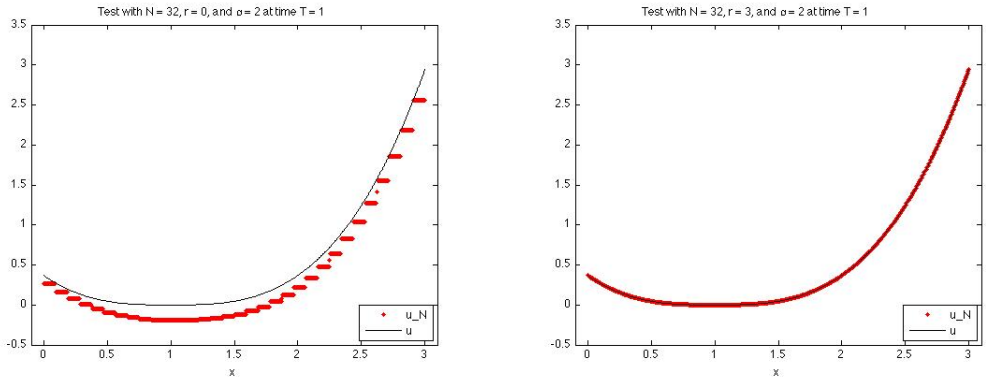


(b) $r = 3$ with $\kappa_t = 0.0005$.

Figure 3.27: Computed solutions at time $t = 1$ for Example 3.12 using RK4 time-stepping with $\alpha = 2$, $h = 9.375\text{e-}02$, $\delta = \frac{1}{h}$, and $u_h^0 = \mathcal{P}_h u_0$.

Table 3.24: Rates of convergence in space for Example 3.12 at time $t = 1$ using trapezoidal time-stepping with $\alpha = 2$, $\Delta t = 0.001$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = 3/4$	$h = 3/8$		$h = 3/16$		$h = 3/32$	
		Error	Error	Order	Error	Order	Error	Order
0	L^2	8.6e-01	8.8e-01	-0.02	5.9e-01	0.58	2.9e-01	1.00
	L^∞	1.8e+00	1.3e+00	0.53	7.3e-01	0.80	3.9e-01	0.92
1	L^2	2.0e-01	7.3e-02	1.45	1.3e-02	2.44	3.2e-03	2.05
	L^∞	3.9e-01	1.0e-01	1.92	2.7e-02	1.95	6.8e-03	1.98
2	L^2	1.1e-01	2.0e-02	2.49	5.2e-03	1.98	1.3e-03	2.02
	L^∞	1.1e-01	2.0e-02	2.42	5.2e-03	1.99	1.3e-03	2.00
3	L^2	3.0e-02	6.8e-03	2.13	1.7e-03	2.00	4.2e-04	2.03
	L^∞	2.9e-02	6.9e-03	2.08	1.7e-03	2.02	4.2e-04	2.02



(a) $r = 0$. **(b)** $r = 3$.

Figure 3.28: Computed solutions at time $t = 1$ for Example 3.12 using trapezoidal time-stepping with $\alpha = 2$, $h = 1.963\text{e-}01$, $\Delta t = 0.001$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

Table 3.25: Rates of convergence in space for Example 3.13 at time $t = 1$ using backwards Euler time-stepping with $r = 0$, $\alpha = 2I$, $\Delta t = 0.1$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

h	L^∞ norm	order	L^2 norm	order
2.83e-01	5.62e-01		2.63e-01	
1.77e-01	3.62e-01	0.93	1.71e-01	0.92
1.41e-01	2.92e-01	0.96	1.38e-01	0.96

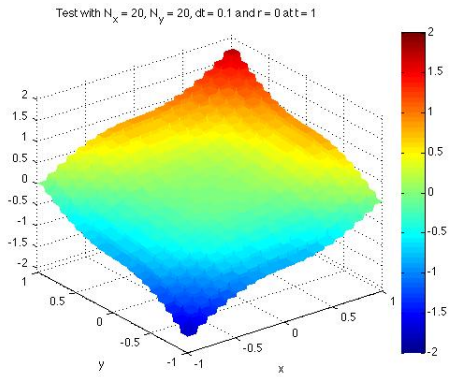


Figure 3.29: Computed solution at time $t = 1$ for Example 3.13 using backwards Euler time-stepping with $r = 0$, $\alpha = 2I$, $h = 1.414\text{e-}01$, $\Delta t = 0.1$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

Table 3.26: Rates of convergence in space for Example 3.13 at time $t = 1$ using backwards Euler time-stepping with $r = 1$, $\alpha = 2I$, $\Delta t = 0.1$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

h	L^∞ norm	order	L^2 norm	order
4.71e-01	7.41e-02		5.00e-02	
3.54e-01	4.21e-02	1.96	3.56e-02	1.18
2.83e-01	3.10e-02	1.38	2.76e-02	1.14

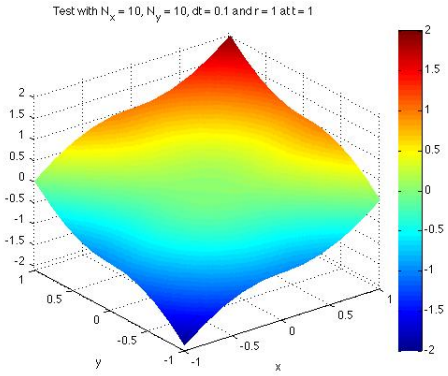


Figure 3.30: Computed solution at time $t = 1$ for Example 3.13 using backwards Euler time-stepping with $r = 1$, $\alpha = 2I$, $h =$ 2.828e-01, $\Delta t = 0.1$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

Table 3.27: Rates of convergence in time for Example 3.13 at time $t = 1$ using backwards Euler time-stepping with $r = 2$, $\alpha = 2I$, $h = 1.414$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

Δt	L^∞ norm	order	L^2 norm	order
5.00e-01	4.12e-02		4.12e-02	
2.50e-01	2.11e-02	0.96	2.11e-02	0.97
1.00e-01	8.55e-03	0.99	8.49e-03	0.99
5.00e-02	4.29e-03	1.00	4.25e-03	1.00

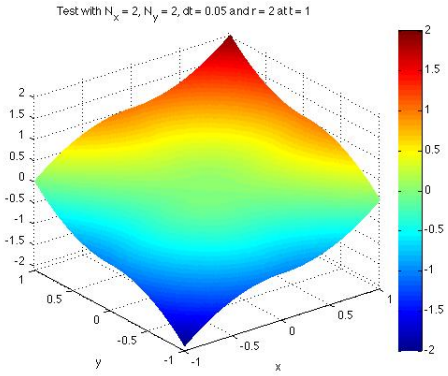


Figure 3.31: Computed solution at time $t = 1$ for Example 3.13 using backwards Euler time-stepping with $r = 2$, $\alpha = 2I$, $h = 1.414$, $\Delta t = 0.05$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

Chapter 4

Interior Penalty Discontinuous Galerkin Methods

In this chapter, we propose another class of DG methods for directly approximating the viscosity solutions of fully nonlinear second order elliptic and parabolic PDE problems using an interior penalty discontinuous Galerkin (IPDG) methodology. Our formulation will now be targeted towards approximating viscosity solutions with C^1 regularity, although we will see in Section 4.4 that the methods appear to work for viscosity solutions with less regularity. The methods will be based on a nonstandard mixed formulation. The main idea will be to introduce various Hessian approximations using an IPDG methodology, and then introduce a numerical operator that incorporates the various discrete Hessians in forming a numerical moment. Much of the motivation and formulation follows directly from the monotone DG framework developed in Section 3.2, where a nonstandard LDG methodology was utilized. The main difference will be the fact that we do not form multiple approximations for ∇u because of the assumed C^1 regularity. As such, we do not enforce a g-monotonicity requirement with regards to first-order terms. The methods developed in this chapter heavily rely upon the DG notation developed in Section 3.1. We also adopt the

convention in Section 3.2.4 for higher-order bases by letting all boundary values be given by interior limits.

4.1 A Monotone Framework for Second Order Elliptic Problems

We now develop a class of IPDG methods for second order boundary value problems of the form

$$F[u](x) := F\left(D^2u, \nabla u, u, x\right) = 0, \qquad x \in \Omega \subset \mathbb{R}^d, \tag{4.1a}$$

$$u(x) = g(x), \qquad x \in \partial\Omega, \tag{4.1b}$$

where F is a fully nonlinear elliptic operator, Ω is an open, bounded, convex domain, and the viscosity solution $u \in C^1(\Omega)$. We will see in the numerical tests found in Section 4.4 that the framework also appears to be well-suited for approximating conditionally elliptic problems with viscosity solutions that have only H^1 regularity.

4.1.1 Motivation

Due to the lower regularity of the viscosity solution, D^2u may not exist. Thus, we will use three possible approximations for D^2u, namely, the left and right limits, as well as their average. Using a numerical operator that can handle multiple approximations for D^2u, we rewrite (4.1) in mixed form as

$$\widehat{F}(P_1, P_2, P_3, \nabla u, u, x) = 0, \tag{4.2a}$$

$$P_1(x) - D^2u(x^+) = 0, \tag{4.2b}$$

$$P_2(x) - D^2u(x^a) = 0, \tag{4.2c}$$

$$P_3(x) - D^2u(x^-) = 0 \tag{4.2d}$$

for all $x \in \Omega$, where $D^2u(x^a)$ can be thought of as the arithmetic average of $D^2u(x^+)$ and $D^2u(x^-)$.

We now formalize the definition and properties of a numerical operator. The following definitions follow directly from their counterparts in Section 2.5.

Definition 4.1.

(i) *The function* $\widehat{F} : \left(\mathbb{R}^{d\times d}\right)^3 \times \mathbb{R}^d \times \mathbb{R} \times \Omega \to \mathbb{R}$ *in* (4.2a) *is called a* numerical operator.

(ii) *Let* $P \in \overline{\mathbb{R}}^{d\times d}$, $q \in \overline{\mathbb{R}}^d$, $v \in \mathbb{R}$, *and* $x \in \overline{\Omega}$. *The numerical operator* $\widehat{F}$ *in* (4.2a) *is said to be* consistent *if* $\widehat{F}$ *satisfies*

$$\liminf_{\substack{P_1,P_2,P_3\to P,\zeta\to q;\\ \nu\to v,\ \xi\to x}} \widehat{F}(P_1,P_2,P_3,\zeta,\nu,\xi) \geq F_*(P,q,v,x),$$
$$\limsup_{\substack{P_1,P_2,P_3\to P,\zeta\to q;\\ \nu\to v,\ \xi\to x}} \widehat{F}(P_1,P_2,P_3,\zeta,\nu,\xi) \leq F^*(P,q,v,x),$$

where F_* *and* F^* *denote, respectively, the lower and the upper semi-continuous envelopes of* F. *Thus, we have*

$$F_*(P,q,v,x) := \liminf_{\substack{\widetilde{P}\to P,\widetilde{q}\to q,\\ \widetilde{v}\to v,\widetilde{x}\to x}} F\left(\widetilde{P},\widetilde{q},\widetilde{v},\widetilde{x}\right),$$
$$F^*(P,q,v,x) := \limsup_{\substack{\widetilde{P}\to P,\widetilde{q}\to q,\\ \widetilde{v}\to v,\widetilde{x}\to x}} F\left(\widetilde{P},\widetilde{q},\widetilde{v},\widetilde{x}\right),$$

where $\widetilde{P} \in \mathbb{R}^{d\times d}$, $\widetilde{q} \in \mathbb{R}^d$, $\widetilde{v} \in \mathbb{R}$, *and* $\widetilde{x} \in \Omega$.

(iii) *The numerical operator* $\widehat{F}$ *in* (4.2a) *is said to be* generalized monotone (g-monotone) *if* $\widehat{F}(P_1,P_2,P_3,q,v,x)$ *is monotone increasing in* P_1 *and* P_3 *and monotone decreasing in* P_2. *More precisely, for all* $P_1,P_2,P_3 \in \mathbb{R}^{d\times d}$, $q \in \mathbb{R}^d$,

$v \in \mathbb{R}$, *and* $x \in \Omega$, *there holds*

$$\begin{aligned}
\widehat{F}\left(A, P_2, P_3, q, v, x\right) &\leq \widehat{F}\left(B, P_2, P_3, q, v, x\right), \\
\widehat{F}\left(P_1, A, P_3, q, v, x\right) &\geq \widehat{F}\left(P_1, B, P_3, q, v, x\right), \\
\widehat{F}\left(P_1, P_2, A, q, v, x\right) &\leq \widehat{F}\left(P_1, P_2, B, q, v, x\right),
\end{aligned}$$

for all $A, B \in \mathcal{S}^{d\times d}$ *such that* $A \leq B$, *where* $A \leq B$ *means that* $B - A$ *is a nonnegative definite matrix. In other words,* $\widehat{F}(\uparrow, \downarrow, \uparrow, q, v, x)$.

In order to ensure the g-monotonicity property of the numerical operator in (4.2a), we introduce a *numerical moment*. Then, we have the following two examples of Lax-Friedrichs-like numerical operators that have been modified for the IPDG framework:

$$\widehat{F}_1(P_1, P_2, P_3, q, \lambda, \xi) := F(P_2, q, \lambda, \xi) + \alpha_1 : \big(P_1 - 2P_2 + P_3\big), \tag{4.4a}$$

$$\widehat{F}_2(P_1, P_2, P_3, q, \lambda, \xi) := F\Big(\frac{P_1 + P_2 + P_3}{3}, q, \lambda, \xi\Big) + \alpha_2 : \big(P_1 - 2P_2 + P_3\big), \tag{4.4b}$$

where $\alpha_1, \alpha_2 \in \mathbb{R}^{d\times d}$ are positive constant matrices chosen to enforce the g-monotonicity property of $\widehat{F}$ and the last term in (4.4a) and (4.4b) is called the numerical moment. To ensure $\widehat{F}_1$ is g-monotone, we require

$$\alpha_1 > \pm\frac{\partial F}{\partial D^2 u}, \tag{4.5}$$

assuming adequate regularity for the differential operator F.

4.1.2 Formulation

We now formalize our IPDG discretization of (4.2). As in Section 3.2.2, we discretize (4.2a) by simply using its L^2-projection into V^h, namely,

$$a_0\big(u_h, P_{1h}, P_{2h}, P_{3h}; \phi_{0h}\big) = 0 \qquad \forall \phi_{0h} \in V^h, \tag{4.6}$$

where

$$a_0\big(u, P_1, P_2, P_3; \phi_0\big) = \Big(\widehat{F}(P_1, P_2, P_3, \nabla u, u, \cdot), \phi_0\Big)_{\mathcal{T}_h}.$$

Note, for $u_h \in V^h$, ∇u_h is defined locally on each simplex $K \in \mathcal{T}_h$.

Next, we discretize the linear auxiliary equations in (4.2). To this end, we first introduce some well-known identities (often referred to as "magic formulas" in the literature) defined on $\mathcal{E}_h^I$ for functions in V^h:

$$\big[v\, w\big] = T_\ell^-(v)\big[w\big] + \big[v\big]T_\ell^+(w), \tag{4.7a}$$

$$\big[v\, w\big] = \big\{v\big\}\big[w\big] + \big[v\big]\big\{w\big\}, \tag{4.7b}$$

$$\big[v\, w\big] = T_\ell^+(v)\big[w\big] + \big[v\big]T_\ell^-(w) \tag{4.7c}$$

on $\mathcal{E}_h^I$ for all $v, w \in V^h$, for all $\ell = 1, 2, \ldots, d$. The flux operators $T^\pm$ are defined locally in Section 3.2.2.

We also introduce interior penalty terms. Let $\gamma_{k,\ell}^{0i} \geq 0$ for $i = 1, 2, 3$ and $k, \ell = 1, 2, \ldots, d$ denote interior penalty parameters. It will be clear later that to avoid redundancy of the three equations for P_{1h}, P_{2h}, and P_{3h}, we need to require that $\gamma_{k,\ell}^{02} \geq \max\{\gamma_{k,\ell}^{01}, \gamma_{k,\ell}^{03}\}$ for all $k, \ell = 1, 2, \ldots, d$ and $\gamma_{k,k}^{02} > \max\{\gamma_{k,k}^{01}, \gamma_{k,k}^{03}\}$ for all $k = 1, 2, \ldots, d$. Then, we define the interior penalty terms by

$$J_{k,\ell}^{0i}(v, w) := \sum_{e \in \mathcal{E}_h^I} \frac{\gamma_{k,\ell}^{0i}}{h_e} \Big\langle \big[v\big], \big[w\big] \Big\rangle_e, \tag{4.8}$$

where

$$h_e := \begin{cases} \operatorname{diam} e, & \text{if } d \geq 2 \\ \max\big\{\operatorname{diam} K_1, \operatorname{diam} K_2 \mid K_1, K_2 \in \mathcal{T}_h \text{ with } \overline{K}_1 \cap \overline{K}_2 = e\big\}, & \text{if } d = 1. \end{cases}$$

In order to discretize the auxiliary linear equations in (4.2), we use the integration by parts formula

$$\int_S v_{x_k x_\ell}\,\varphi\,dx = \int_{\partial S} v_{x_k}\,\varphi\,n_\ell\,ds - \int_S v_{x_k}\,\varphi_{x_\ell}\,dx$$

for all $v, \varphi \in H^2(S)$, for $k, \ell = 1, 2, \ldots, d$. Letting $P_{k,\ell}$ be an approximation for $u_{x_k x_\ell}$, we formally define $P_{k,\ell}$ by

$$\Big(P_{k,\ell}, \phi\Big)_{\mathcal{T}_h} = \Big\langle \big[u_{x_k}\,\phi\big], n_\ell \Big\rangle_{\mathcal{E}_h^I} + \Big\langle u_{x_k}, \phi\, n_\ell \Big\rangle_{\mathcal{E}_h^B} - \Big(u_{x_k}, \phi_{x_\ell}\Big)_{\mathcal{T}_h} \qquad \forall \phi \in H^1(\mathcal{T}_h) \quad (4.9)$$

for all $u \in H^1(\mathcal{T}_h)$, for $k, \ell = 1, 2, \ldots, d$.

Combining the integral identity (4.9) with (4.7) and introducing standard penalty and jump terms, we can now fully discretize the auxiliary linear equations in (4.2). To this end, we let $\epsilon_{k,\ell}^i \in \{-1, 0, 1\}$, $i = 1, 2, 3$ and $k, \ell = 1, 2, \ldots, d$, denote the "symmetrization" parameters. Then, we define the auxiliary variables $P_{1h}, P_{2h}, P_{3h} \in \big[V^h\big]^{d\times d}$ by

$$\Big(P_{k,\ell}^{ih}, \phi_{k,\ell}^{ih}\Big)_{\mathcal{T}_h} + a_{k,\ell}^i\left(u_h, \phi_{k,\ell}^{ih}\right) = f_{k,\ell}^i\left(\phi_{k,\ell}^{ih}\right) \qquad \forall \phi_{k,\ell}^{ih} \in V^h \qquad (4.10)$$

for all $i = 1, 2, 3$ for all $k, \ell = 1, 2, \ldots, d$, where

$$a_{k,\ell}^1(u,\phi) = b_{k,\ell}^1(u,\phi) - \Big\langle T_\ell^-(u_{x_k}), \big[\phi\big] n_\ell \Big\rangle_{\mathcal{E}_h^I} + \epsilon_{k,\ell}^1 \Big\langle \big[u\big], T_\ell^-(\phi_{x_k}) n_\ell \Big\rangle_{\mathcal{E}_h^I}, \qquad (4.11a)$$

$$a_{k,\ell}^2(u,\phi) = b_{k,\ell}^2(u,\phi) - \Big\langle \big\{u_{x_k}\big\}, \big[\phi\big] n_\ell \Big\rangle_{\mathcal{E}_h^I} + \epsilon_{k,\ell}^2 \Big\langle \big[u\big], \big\{\phi_{x_k}\big\} n_\ell \Big\rangle_{\mathcal{E}_h^I}, \qquad (4.11b)$$

$$a_{k,\ell}^3(u,\phi) = b_{k,\ell}^3(u,\phi) - \Big\langle T_\ell^+(u_{x_k}), \big[\phi\big] n_\ell \Big\rangle_{\mathcal{E}_h^I} + \epsilon_{k,\ell}^3 \Big\langle \big[u\big], T_\ell^+(\phi_{x_k}) n_\ell \Big\rangle_{\mathcal{E}_h^I}, \qquad (4.11c)$$

$$f_{k,\ell}^i(\phi) = \epsilon_{k,\ell}^i \sum_{e\in\mathcal{E}_H^B} \Big\langle g, \phi_{x_k} n_\ell \Big\rangle_{\widetilde{e}} + \sum_{e\in\mathcal{E}_h^B} \frac{\gamma_{k,\ell}^{0i}}{h_e} \Big\langle g, \phi \Big\rangle_{\widetilde{e}}, \qquad (4.11d)$$

and $b^i_{k,\ell} : H^1(\mathcal{T}_h) \times H^1(\mathcal{T}_h) \to \mathbb{R}$ is defined by

$$b^i_{k,\ell}(v,w) = \Big(v_{x_k}, w_{x_\ell}\Big)_{\mathcal{T}_h} - \Big\langle v_{x_k}, w\, n_\ell \Big\rangle_{\mathcal{E}_h^B} + \epsilon^i_{k,\ell} \Big\langle v, w_{x_k}\, n_\ell \Big\rangle_{\mathcal{E}_h^B} + J^{0i}_{k,\ell}(v,w) + \sum_{e\in\mathcal{E}_h^B} \frac{\gamma^{0i}_{k,\ell}}{h_e} \Big\langle v, w \Big\rangle_e$$

for all $v, w \in H^1(\mathcal{T}_h)$, for $P^{ih}_{k,\ell} = [P_{ih}]_{k,\ell}$, $i = 1,2,3$ and $k,\ell = 1,2,\ldots,d$.

Suppose $r \geq 1$ in the definition of V^h. Then, our IPDG methods for the fully nonlinear Dirichlet problem (4.1) is defined as seeking $u_h \in V^h$ and $P_{1h}, P_{2h}, P_{3h} \in \big[V^h\big]^{d\times d}$ such that (4.6) and (4.10) hold for all $i = 1,2,3$ and $k,\ell = 1,2,\ldots,d$.

4.1.3 The Numerical Moment

We now take a closer look at the numerical moment used in the definition of the Lax-Friedrichs-like numerical operator, (4.4a). For simplicity, we assume the three symmetrization constants are the same, i.e., $\epsilon^1 = \epsilon^2 = \epsilon^3$. Then,

$$\begin{aligned}
\Big(P^{1h}_{k,\ell} - 2P^{2h}_{k,\ell} + P^{3h}_{k,\ell}, \phi\Big)_{\mathcal{T}_h} &= f^1_{k,\ell}(\phi) - 2f^2_{k,\ell}(\phi) + f^3_{k,\ell}(\phi) \\
&\quad - a^1_{k,\ell}(u_h,\phi) + 2a^2_{k,\ell}(u_h,\phi) - a^3_{k,\ell}(u_h,\phi) \\
&= \sum_{e\in\mathcal{E}_h^B} \frac{\gamma^{01}_{k,\ell} - 2\gamma^{02}_{k,\ell} + \gamma^{03}_{k,\ell}}{h_e} \Big(\Big\langle g, \phi(x^I) \Big\rangle_{\widetilde{e}} - \Big\langle u_h, \phi \Big\rangle_e \Big) \\
&\quad - \sum_{e\in\mathcal{E}_h^I} \frac{\gamma^{01}_{k,\ell} - 2\gamma^{02}_{k,\ell} + \gamma^{03}_{k,\ell}}{h_e} \Big\langle [u_h], [\phi] \Big\rangle_e
\end{aligned}$$

for all $\phi \in V^h$. Thus, we have

$$
\begin{aligned}
&a_0\big(u_h, P_{1h}, P_{2h}, P_{3h}; \phi\big) \\
&\quad = \Big(F\left(P_{2h}, \nabla u_h, u_h, \cdot\right), \phi\Big)_{\mathcal{T}_h} + \sum_{e \in \mathcal{E}_h^I} \frac{\alpha_1 : (2\gamma^{02} - \gamma^{01} - \gamma^{03})}{h_e} \Big\langle \big[u_h\big], \big[\phi\big] \Big\rangle_e \\
&\qquad + \sum_{e \in \mathcal{E}_h^B} \frac{\alpha_1 : (2\gamma^{02} - \gamma^{01} - \gamma^{03})}{h_e} \Big(\Big\langle u_h, \phi \Big\rangle_e - \Big\langle g, \phi \Big\rangle_{\widetilde{e}} \Big),
\end{aligned}
$$

and it follows that our IPDG discretization amounts to replacing the continuous Hessian operator with a discrete Hessian operator, projecting our fully nonlinear operator into the DG space, and adding a penalization term to the nonlinear equation.

4.1.4 Remarks about the Formulation

We end this section with a few remarks.

Remark 4.1.

(a) Looking backwards, (4.10) provides a proper interpretation for each auxiliary variable P_{1h}, P_{2h}, and P_{3h} for a given function u_h. Each P_{ih} defines a discrete Hessian for u_h. The functions P_{1h}, P_{2h}, and P_{3h} should be very close to each other if D^2u exists and is continuous. However, their discrepancies are expected to be large if D^2u does not exist.

(b) We have the three equations for approximating $u_{x_k x_\ell}$, i.e., (4.10) for $k, \ell \in \{1, 2, \ldots, d\}$ fixed and $i = 1, 2, 3$, are linearly independent provided that $\gamma_{k,\ell}^{02} > \max\{\gamma_{k,\ell}^{01}, \gamma_{k,\ell}^{03}\}$.

(c) Comparing the jump formulations found in Sections 4.1.3 and 3.2.6, we see that the numerical moment term used in the IPDG methodology is actually analogous to the numerical viscosity term used in the LDG methodology. The exact numerical moment term from Chapter 3 is based upon using two approximations for ∇u. However, by Remark 2.4, we see that the numerical viscosity is an

essential tool in the FD convergence proof. We will see in Section 4.4 that the proposed IPDG methods still perform well for many numerical test problems, especially when paired with an appropriate nonlinear solver.

(d) *We could also add a high-order interior-penalty term to* $\widehat{F}$ *such as*

$$J^{1i}_{k,\ell}(v,w) := \sum_{e\in\mathcal{E}_h^I} \frac{\gamma^{1i}_{k,\ell}}{h_e}\Big\langle \big[\nabla v\big], \big[\nabla w\big]\Big\rangle_e,$$

where $\gamma^{1i}_{k,\ell} \geq 0$ *for all* $i = 1,2,3$, $k,\ell = 1,2,\ldots,d$. *However, such a term is not analogous to the jump formulation of the numerical moment used in Chapter 3. Also, such a term would not be appropriate if approximating a viscosity solution that is only continuous or* H^1. *We will see in Section 4.4 that the proposed IPDG methods can be used to approximate viscosity solutions that have only* H^1 *regularity.*

(e) *The reason for requiring* $r \geq 1$ *can be explained as follows. When* $r = 0$, *the piecewise constant functions have piecewise zero derivatives on the given mesh. After eliminating the jump terms containing derivatives in (4.10), it is clear that the ability for* P_{1h} *and* P_{3h} *to carry information from the left and the right, respectively, is lost. Furthermore, if* $\gamma^{01}_{k,\ell} = \gamma^{02}_{k,\ell} = \gamma^{03}_{k,\ell}$, *then* $a^1_{k,\ell} = a^2_{k,\ell} = a^3_{k,\ell}$ *for all* $k,\ell = 1,2,\ldots,d$. *As a result, the numerical moment term vanishes and we are left with a trivial discretization for (4.1), which is known not to work well in general.*

(f) *Let* $\Lambda^i_h := \sum_{k=1}^d [P_{ih}]_{k,k}$ *for* $i = 1,2,3$. *Then, we have*

$$\sum_{k=1}^d a^i_{k,k}\left(u_h, \phi^i_h\right) = \sum_{k=1}^d f^i_{k,k}\left(\phi^i_h\right) - \left(\Lambda^i_h, \phi^i_h\right)_{\mathcal{T}_h} \tag{4.12}$$

for all $\phi^i_h \in V^h$ *for* $i = 1,2,3$. *Treating* $\{\Lambda^i_h\}_{i=1}^3$ *as "sources", (4.12) represents three different Poisson discretizations for* u. *Thus, for* γ_{0i} *sufficiently large,*

we have (4.12) *forms an invertible linear mapping between* Λ_h^i *and* u_h *for all* $i = 1, 2, 3$. *Furthermore, the mapping is symmetric for a given* $i \in \{1, 2, 3\}$ *if* $\epsilon_{k,k}^i = -1$ *for all* $k = 1, 2, \ldots, d$. *We call the mapping "nonsymmetric" if* $\epsilon_{k,k}^i = 1$ *for all* $k = 1, 2, \ldots, d$, *and we call the mapping "incomplete" if* $\epsilon_{k,k}^i = 1$ *for all* $k = 1, 2, \ldots, d$.

(g) Notice that (4.6) *and* (4.10) *forms a nonlinear system of equations, with the nonlinearity only appearing in* a_0. *Thus, a nonlinear solver is necessary in implementing the above scheme. Based on* (4.12), *we can form a solver analogous to Algorithm 3.1. In Section 4.4 we will perform numerical tests using both a straight-forward Newton solver on the entire system and a solver to be proposed that is analogous to Algorithm 3.1. We will see that our proposed discretizations once again either remove or destabilize many of the numerical artifacts that plague a trivial discretization of a fully nonlinear PDE problem.*

(h) Due to the fully nonlinear structure of the PDE, no integration by parts can be performed. As a result, the "primal" form of the nonstandard LDG methods in Chapter 3 is intrinsically different from the "primal" form of the above nonstandard IPDG methods.

4.2 Extensions of the IPDG Framework for Second Order Parabolic Problems

Using the above IPDG methodology for elliptic problems, we now develop a class of fully discrete methods for second order initial-boundary value problems of the form

$$u_t + F\left(D^2 u, \nabla u, u, x, t\right) = 0, \qquad (x,t) \in \Omega_T := \Omega \times (0, T], \tag{4.13a}$$

$$u(x,t) = g(x,t), \qquad (x,t) \in \partial\Omega \times (0, T], \tag{4.13b}$$

$$u(x,0) = u_0(x), \qquad x \in \Omega, \tag{4.13c}$$

where F is a fully nonlinear elliptic operator, Ω is an open, bounded, convex domain, T is a positive number, and the viscosity solution $u \in C^1\big(\Omega \times (0,T]\big)$. The methodology will be based on using the method of lines for the time discretization.

To partition the time domain, we fix an integer $M > 0$ and let $\Delta t = \frac{T}{M}$. Then, we define $t_k := k\,\Delta t$ for $0 \leq k \leq M$. Notationally, $u_h^k(x) \in V^h$ and $P_{ih}^k \in \left[V^h\right]^{d\times d}$ will be approximations for $u(x,t_k)$ and $D^2u(x,t_k)$, respectively, for all $0 \leq k \leq M$. For both implicit and explicit schemes, we define the initial value, u_h^0, by

$$u_h^0 = \mathcal{P}_h u_0, \tag{4.14}$$

where the projection operator $\mathcal{P}_h$ is defined by (3.4).

To simplify the appearance of the methods and to make them more transparent for use with a given ODE solver, we define discrete Hessian operators $D_{ih}^{2,k} : V^h \to \left[V^h\right]^{d\times d}$ for $i = 1,2,3$ at time t_k using (4.10), where $0 \leq k \leq M$. Then, we define $D_{ih}^{2,k}v$ by

$$\begin{aligned}\Big(\Big[D_{ih}^{2,k}v\Big]_{\ell,m}, \phi_{\ell,m}^{ih}\Big)_{\mathcal{T}_h} + a_{\ell,m}^i\left(v,\phi_{\ell,m}^{ih}\right) \\ = \epsilon_{\ell,m}^i\Big\langle g(\cdot,t_k), \left(\phi_{\ell,m}^{ih}\right)_{x_\ell} n_m\Big\rangle_{\partial\Omega} + \sum_{e\in\mathcal{E}_h^B}\frac{\gamma_{\ell,m}^{0i}}{h_e}\Big\langle g(\cdot,t_k),\phi_{\ell,m}^{ih}\Big\rangle_{\widetilde{e}} \qquad \forall \phi_{\ell,m}^{ih} \in V^h.\end{aligned} \tag{4.15}$$

We also introduce the operator notation

$$\widehat{F}^k\left[v\right] := \widehat{F}\left(D_{1h}^{2,k}v, D_{2h}^{2,k}v, D_{3h}^{2,k}v, \nabla v, v, x, k\Delta t\right) \tag{4.16}$$

for all $v \in V^h$. Then, we have the semi-discrete equation

$$\frac{\partial}{\partial t}u_h(x,t_k) = -\mathcal{P}_h\widehat{F}^k\big[u_h(x,t_k)\big] \tag{4.17}$$

for all $0 \leq k \leq M$, $x \in \Omega$.

Letting $\mathcal{P}_{h,k}$ denote the modified projection operator defined by (3.38), we can define fully discrete methods for approximating problem (4.13) based on approximating (4.17) using the forward Euler method, backward Euler method, and the trapezoidal method. Hence, we have the following fully discrete schemes for approximating (4.13):

$$u_h^{n+1} = \mathcal{P}_{h,n+1}\left(u_h^n - \Delta t\, \widehat{F}^n\left[u_h^n\right]\right), \tag{4.18}$$

$$u_h^{n+1} + \Delta t\, \mathcal{P}_h\, \widehat{F}^{n+1}\left[u_h^{n+1}\right] = u_h^n, \tag{4.19}$$

and

$$u_h^{n+1} + \frac{\Delta t}{2}\, \mathcal{P}_h\, \widehat{F}^{n+1}\left[u_h^{n+1}\right] = u_h^n - \frac{\Delta t}{2}\, \mathcal{P}_h\, \widehat{F}^n\left[u_h^n\right] \tag{4.20}$$

for $n = 0, 1, \ldots, M-1$, where $u_h^0 := \mathcal{P}_h u_0$ and, for (4.19) and (4.20), we also have the implied equations $P_{ih}^n = D_{ih}^{2,n} u_h^n$ for all $i = 1, 2, 3$. Observe, (4.18), (4.19), and (4.20) correspond to the forward Euler method, backward Euler method, and trapezoidal method, respectively.

We can also formulate RK methods for approximating (4.17) as follows. Let s be a positive integer, $A \in \mathbb{R}^{s\times s}$, and $b, c \in \mathbb{R}^s$ such that

$$\sum_{\ell=1}^{s} a_{k,\ell} = c_k$$

for each $k = 1, 2, \ldots, s$. Then, a generic s-stage RK method for approximating (4.17) can be written

$$u_h^{n+1} = \mathcal{P}_{h,n+1}\left(u_h^n - \Delta t \sum_{\ell=1}^{s} b_\ell \widehat{F}^{n+c_\ell}[\xi_h^{n,\ell}]\right) \tag{4.21}$$

with

$$\xi_h^{n,\ell} = \mathcal{P}_{h,n+c_k}\left(u_h^n - \Delta t \sum_{k=1}^{s} a_{k,\ell} \widehat{F}^{n+c_k}[\xi_h^{n,k}]\right)$$

for all $n = 0, 1, \ldots, N-1$ and $u_h^0 = \mathcal{P}_h u_0$. We note that (4.21) corresponds to an explicit method when A is strictly lower diagonal and an implicit method otherwise.

Also, we can interpret $\xi_h^{n,\ell}$ in (4.21) as an approximation for $u_h^{n+c_\ell}$. Since the boundary condition at time t_{n+1} is enforced by $\widehat{F}^{n+1}$, we can set $\delta = 0$ in (3.38) if $c_s = 1$.

4.3 General Solvers

In this section, we discuss the different strategies from Section 3.4 adapted for solving the nonlinear system of equations that results from the IPDG discretization for the elliptic problem, (4.1). Again, we have a nonlinear equation that is complemented by an auxiliary system of linear equations. Furthermore, the nonlinear equation is monotone in three of its five arguments at every given point in the domain. Many of the numerical tests in Section 4.4 will simply use a nonlinear solver for the full system of equations. However, in this section, we propose algorithms for two solvers that have been tailored towards the IPDG discretization and discuss the benefits of the two different algorithms in Remark 4.2.

Our first observation is that the numerical operators given by (4.4) are symmetric in P_{1h} and P_{3h}. Thus, as before, there is the possibility for variable reduction. Assuming $\gamma_{k,\ell}^{01} = \gamma_{k,\ell}^{03}$ for all $k, \ell = 1, 2, \ldots, d$, we can form a new variable $P_h = \frac{P_{1h}+P_{3h}}{2} \in V^h$. Then, we have P_h and P_{2h} correspond to two IPDG approximations for the Hessian of u that both use averaged flux values on all interior faces/edges, i.e., P_h and P_{2h} are both solutions to (4.10) using $i = 2$ where the only difference is the value of the penalty parameter γ. Let D_{ih}^2 denote the discrete Hessian operator $D_{ih}^{2,k}$ defined by (4.15) with the time dependence k removed. Then, we have the discrete Hessian operators D_{2h}^2 and $D_h^2 = \frac{D_{1h}^2+D_{3h}^2}{2}$ can be considered analogous to the two "centered" discrete Hessian operators $\overline{D}_h^2$ and $\widetilde{D}_h^2$ defined in Section 3.4.2. Using the above observation, we immediately can form the direct solver for the reduced system of equations using $\widehat{F}_1$ defined by (4.4a):

Algorithm 4.1.

1. *Given $\mathcal{T}_h$ and V^h, compute the operators D_h^2 and D_{2h}^2.*

2. *Solve the single nonlinear equation*

$$0 = \left(F\left(D_{2h}^2 u_h, \nabla u_h, u_h, \cdot\right) + 2\alpha : \left(D_h^2 u_h - D_{2h}^2 u_h\right), \phi_h \right)_{\mathcal{T}_h} \qquad \forall \phi_h \in V^h$$

for $u_h \in V^h$.

We now present a splitting algorithm that relies upon the observation in Remark 4.1 part (d) and the invertibility of (4.12). The algorithm is based on using the numerical moment as a means to split the system of equations.

Algorithm 4.2.

1. *Pick initial guesses for u_h, P_{1h}, and P_{3h}.*

2. *Set*

$$\left[G\right]_\ell := F\left(P_{2h}, \nabla u_h, u_h, x\right) + \widehat{\alpha}\left[P_{1h} - 2P_{2h} + P_{3h}\right]_{\ell,\ell}$$

for a fixed constant $\widehat{\alpha} > 0$, and solve

$$\left(\left[G\right]_\ell, \phi_\ell \right)_{\mathcal{T}_h} = 0 \qquad \forall \phi_\ell \in V^h$$

for $[P_{2h}]_{\ell,\ell}$ for all $\ell = 1, 2, \ldots, d$.

3. *Set $\Lambda_h^2 = \sum_{\ell=1}^{d} [P_{2h}]_{\ell,\ell}$. Find u_h by solving (4.12) for the given value of Λ_h^2.*

4. *Set $P_{1h} = D_{1h}^2 u_h$ and $P_{3h} = D_{3h}^2 u_h$.*

5. *Repeat Steps 2 - 4 until the change in P_{2h} is sufficiently small.*

We end the section with a couple of remarks concerning the observed performance of the proposed algorithms.

Remark 4.2.

(a) *Algorithm 4.1 appears to perform faster than using a standard Newton solver on the full system of equations that results from the mixed formulation. This is expected due to the significantly reduced number of unknowns. However, the algorithm appears to be less selective than when the full system of equations is used. We will see some of the benefits of using the full system of equations versus the direct solver in Section 4.4. In contrast, we note that the direct solver for the DG framework in Section 3.2 appears to still preserve the selectivity of the full system of equations from the mixed formulation, especially when using* $r = 0$ *or* $r = 1$*. A potential explanation is the fact that the numerical moment formed by the operators* $\overline{D}_h^2$ *and* $\widetilde{D}_h^2$ *is more selective than the numerical moment formed by the operators* D_h^2 *and* D_{2h}^2*, especially in light of Remark 4.1 part (c).*

(b) *Algorithm 4.2 appears to be more selective than using a standard Newton solver on the full system of equations that results from the mixed formulation. However, the algorithm also appears to converge more slowly than the standard Newton solver when the Newton solver does converge. Thus, the algorithm may be best utilized as a way to precondition an initial guess for a more efficient solver.*

(c) *There is potential to speed up Algorithm 4.2. Step 2 of Algorithm 4.2 requires solving a nonlinear system that is entirely monotone with respect to the unknowns for* $\widehat{\alpha}$ *sufficiently large. Furthermore, the nonlinear equation is entirely local with respect to* $\mathcal{T}_h$*, and can be solved in parallel. Step 3 of Algorithm 4.2 requires inverting a sparse matrix that is symmetric and positive definite when choosing* γ_{kk}^0 *sufficiently large and* $\epsilon_{kk} = -1$ *for all* $k = 1, 2, \ldots, d$*.*

(d) *From Section 4.1.3, we can see that there is a possibility the discretization contains* C^0 *artifacts. Algorithm 4.2 can be interpreted as a fixed-point solver that iterates over the discrete Laplacian. We will see in Section 4.4.3 that by*

iterating over a high-order term in the discretization, Algorithm 4.2 appears to "destabilize" numerical artifacts even when such artifacts are present.

(e) By summing the diagonal elements of the discrete Hessian, we are able to map a second order derivative function in V^h back to u_h. Thus, we will have

$$\sum_{\ell=1}^{d} [P_{2h}]_{\ell\ell} = \sum_{\ell=1}^{d} \left[D_{2h}^2 u_h\right]_{\ell\ell}.$$

However, when u_h is an approximation for a low regularity function, we will not have $[P_{2h}]_{\ell\ell} = [D_{2h}^2 u_h]_{\ell\ell}$ for all $\ell = 1, 2, \ldots, d$. In fact, we would expect inverting the Laplacian operator to have a "smoothing" effect. Therefore, we would have $P_{ih} \neq P_{2h}$ for $i = 1, 3$, and large discrepancies can serve as an indicator for low regularity and/or adaptivity. This observation will be seen in Section 4.4.3 below.

(f) We may not be able to enforce the g-monotonicity requirement globally on a given fully nonlinear PDE such as the Monge-Ampère equation where the differential operator is only elliptic when acting on a particular class of functions. Thus, for such problems, we propose enforcing the g-monotonicity requirement "locally", i.e., over each iteration of the nonlinear solver, as described in the following definition.

Definition 4.2. *The numerical operator $\widehat{F}$ in (4.2a) is said to be* locally generalized monotone (locally g-monotone) *for a function $v_h \in V^h$ if there holds $\widehat{F}\left(D_{1h}^2 v_h, D_{2h}^2 v_h, D_{3h}^2 v_h, \nabla_h v_h, v_h, x\right)$ is monotone increasing in $D_{1h}^2 v_h$ and $D_{3h}^2 v_h$ and monotone decreasing in $D_{2h}^2 v_h$.*

4.4 Numerical Experiments

We use this section to present a series of numerical tests to demonstrate the utility of the proposed IPDG methods for fully nonlinear PDEs of the types (4.1) and (4.13). In

all of our tests we shall use uniform spatial meshes as well as uniform temporal meshes for the dynamic problems. For two-dimensional problems, we use uniform Cartesian partitions on rectangular domains. To solve the resulting nonlinear algebraic systems, we use the Matlab built-in nonlinear solver *fsolve*. When recording the coefficient α for the numerical moment, we let I denote the identity matrix and $\mathbf{1}$ denote the ones matrix for the two-dimensional experiments. For the elliptic problems, we choose the initial guess as the linear interpolant of the boundary data for one-dimensional problems and as the zero function for two-dimensional problems. For dynamic problems, we let $u_h^0 = \mathcal{P}_h u_0$, $P_{1h}^0 = D_{1h}^{2,0} u_h^0$, $P_{2h}^0 = D_{2h}^{2,0} u_h^0$, and $P_{3h}^0 = D_{3h}^{2,0} u_h^0$. Also, the initial guess for u_h^n will be provided by u_h^{n-1}, and the initial guesses for P_{1h}^n, P_{2h}^n, and P_{3h}^n will be provided by P_{1h}^{n-1}, P_{2h}^{n-1}, and P_{3h}^{n-1}, respectively. For convenience, we set $\epsilon^i = 0$, $i = 1, 2, 3$, for all tests except Example 4.3 with $r = 1$. We remark that similar results can be obtained when $\epsilon^i \neq 0$ and the penalty constants are sufficiently large. However, the actual benefit of the symmetrization parameter is unclear in the context of nonlinear algebraic systems. Also, the role of the numerical moment will be further explored in Section 4.4.3.

For our numerical tests, errors will be measured in the L^∞ norm and the L^2 norm, where the errors are measured at the current time step for the dynamic problems. The dynamic test problems will be discretized using both forward and backward Euler methods, as represented by (4.18) and (4.19), respectively. We shall see that the lower order time discretization actually dominates the approximation error for a reasonable time step size Δt when using $r > 1$ in V^h. For the elliptic test problems and for the dynamic test problems where the error is dominated by the spatial discretizations, it appears that the spatial error may have order $\mathcal{O}(h^{\min\{\ell,k\}})$ for $u \in H^k(\Omega)$, where

$$\ell = \begin{cases} r+1, & \text{for } r \text{ odd}, \\ r, & \text{for } r \text{ even}. \end{cases}$$

However, the rates are not perfectly clear from the test data, and there may be some dependence on the regularity of the differential operator F.

4.4.1 Elliptic Problems

We first present the results for six elliptic test problems.

Example 4.1. *Consider the problem*

$$\begin{aligned} -u_{xx}^3 + |u_x| + S(x) &= 0, \qquad -2 < x < 2, \\ u(-2) = \sin(4), \quad u(2) &= -\sin(4), \end{aligned}$$

where

$$S(x) = \left[2\,\mathrm{sign}(x)\cos(x^2) - 4\,x^2\sin(x|x|)\right]^3 - 2|x\cos(x^2)|.$$

This problem has the exact viscosity solution $u(x) = \sin(x|x|) \in H^2(-2,2)$.

Notice that the equation is nonlinear in both u_{xx} and u_x, and the exact solution is not twice differentiable at $x = 0$. The numerical results are shown in Table 4.1 and Figure 4.1. As expected, we can see from the plot that the error appears largest around the node $x = 0$, and both the accuracy and order of convergence improve as the order of the elements increases. We expect that the convergence rates should be bounded by two due to the regularity of the solution. However, for this example, we have $x = 0$ is a node in the partition, and, as a result, we appear to have convergence rates greater than two for $r \geq 3$.

Example 4.2. *Consider the one-dimensional Monge-Ampère problem*

$$\begin{aligned} -u_{xx}^2 + 1 &= 0, \qquad 0 < x < 1, \\ u(0) = 0, \quad u(1) &= 1/2. \end{aligned}$$

Table 4.1: Rates of convergence for Example 4.1 using $\alpha = 4$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $\overline{u}$.

r	Norm	$h = 1$	$h = 1/2$		$h = 1/4$		$h = 1/8$	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	8.1e-01	2.4e-01	1.73	8.0e-02	1.60	2.8e-02	1.52
	L^∞	1.0e+00	2.3e-01	2.14	7.8e-02	1.58	2.7e-02	1.54
2	L^2	1.1e+00	2.9e-01	1.88	4.2e-02	2.78	2.9e-02	0.56
	L^∞	8.1e-01	2.4e-01	1.76	4.5e-02	2.40	1.8e-02	1.30
3	L^2	6.4e-01	2.7e-02	4.55	1.4e-03	4.33	6.5e-05	4.38
	L^∞	4.9e-01	3.1e-02	3.99	1.6e-03	4.32	9.1e-05	4.09
4	L^2	5.6e-02	3.2e-03	4.14	2.4e-04	3.72	1.7e-05	3.83
	L^∞	4.9e-02	3.0e-03	4.02	2.6e-04	3.56	1.6e-05	4.02
5	L^2	2.3e-02	8.5e-04	4.79	1.5e-05	5.82	2.4e-07	5.96
	L^∞	2.1e-02	9.3e-04	4.49	1.8e-05	5.67	2.6e-07	6.11

This problem has exactly two classical solutions

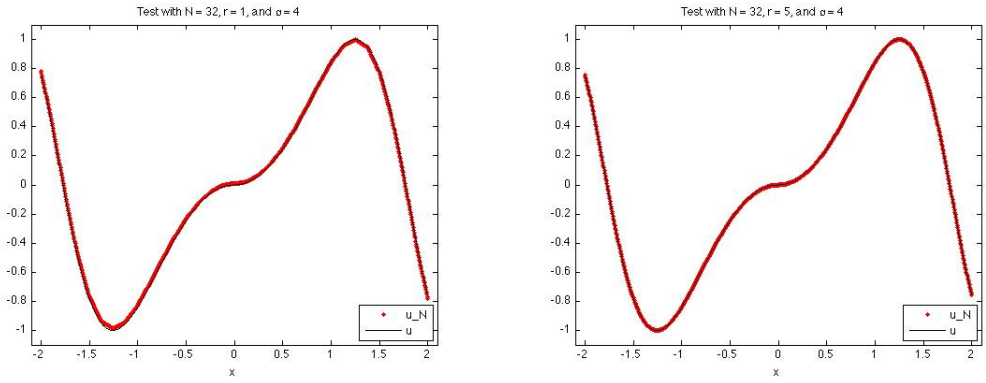
$$u^+(x) = \frac{1}{2}x^2, \qquad u^-(x) = -\frac{1}{2}x^2 + x,$$

where u^+ is convex and u^- is concave. However, u^+ is the unique viscosity solution.

We approximate the given problem using linear elements ($r = 1$) to see how the approximation converges with respect to h when the solution is not in the approximation space V^h. The numerical results are shown in Table 4.2 and Figure 4.2. The results for quadratic elements ($r = 2$) are presented in Table 4.15 and Figure 4.16. We note that the approximations using $r = 2$ are almost exact for each mesh size. This is possible since $u^+ \in V^h$ when $r = 2$.

Example 4.3. *Consider the two-dimensional Monge-Ampère problem*

$$\begin{aligned} -\det D^2 u = -u_{xx}\, u_{yy} + u_{xy}\, u_{yx} &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega, \end{aligned}$$



(a) $r = 1$. **(b)** $r = 5$.

Figure 4.1: Computed solution for Example 4.1 using $\alpha = 4$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, $h =$ 1.250e-01, and *fsolve* with initial guess $\overline{u}$.

Table 4.2: Rates of convergence for Example 4.2 using $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 1$, $\gamma_{02} = 1.1$, $\epsilon^i = 0$, and *fsolve* with initial guess $\overline{u}$.

r	Norm	$h = 1/10$	$h = 1/20$		$h = 1/40$		$h = 1/80$	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	2.9e-03	7.3e-04	2.00	1.8e-04	1.99	4.7e-05	1.97
	L^∞	3.8e-03	9.4e-04	2.00	2.4e-04	1.99	6.1e-05	1.96

where $f = -(1 + x^2 + y^2)e^{x^2+y^2}$, $\Omega = (0, 1) \times (0, 1)$, *and* g *is chosen such that the viscosity solution is given by* $u(x, y) = e^{\frac{x^2+y^2}{2}}$.

We approximate the given problem for $r = 1$ and $r = 2$, with the results recorded in Table 4.3 and Figure 4.3. We observe that the rate of convergence is suboptimal for $r = 1$, and the last approximation showed little improvement even with a refined mesh. Similar results where obtained for $\epsilon^i = 0$ for all $i = 1, 2, 3$. However, for $r = 2$, we observe rates between 2.0 and 2.5, which are superior to the predicted rates of convergence.

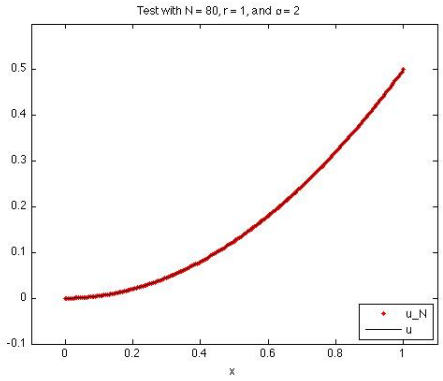


Figure 4.2: Computed solution for Example 4.2 using $r = 1$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 1$, $\gamma_{02} = 1.1$, $\epsilon^i = 0$, $h =$ 2.250e-02, and *fsolve* with initial guess $\overline{u}$.

Table 4.3: Rates of convergence for Example 4.3 using $r = 1$ with $\alpha = 10\,I$, $\gamma_{01} = \gamma_{03} = 100\,\mathbf{1}$, $\gamma_{02} = 200\,\mathbf{1}$, $\epsilon^i = -1\,\mathbf{1}$, and *fsolve* with initial guess $u_h^{(0)} = 0$ and $r = 2$ with $\alpha = 24\,I$, $\gamma_{01} = \gamma_{03} = 20\,\mathbf{1}$, $\gamma_{02} = 40\,\mathbf{1}$, $\epsilon^i = 0$, and 5 iterations of Algorithm 4.2 followed by *fsolve* with initial guess $u_h^{(0)} = 0$.

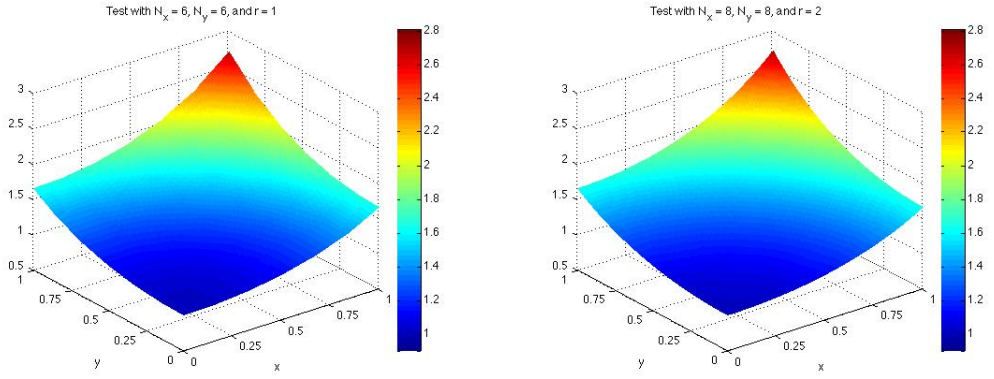
r	Norm	$h =$ 7.07e-01	$h =$ 4.71e-01		$h =$ 3.54e-01		$h =$ 2.83e-01	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	4.88e-02	2.04e-02	2.15	1.31e-02	1.55	1.33e-02	-0.09
	L^∞	1.47e-01	7.82e-02	1.56	4.32e-02	1.65	4.32e-02	0.53
2	L^2	6.37e-03	2.52e-03	2.29	1.30e-03	2.29	7.63e-04	2.39
	L^∞	1.99e-02	7.68e-03	2.35	3.79e-03	2.45	2.17e-03	2.51

Example 4.4. *Consider the two-dimensional Monge-Ampère problem*

$$\begin{aligned} -\det D^2 u = -u_{xx}\,u_{yy} + u_{xy}\,u_{yx} = 0 \qquad & \text{in } \Omega, \\ u = g \qquad & \text{on } \partial\Omega, \end{aligned}$$

where $\Omega = (-1,1) \times (-1,1)$ *and* g *is chosen such that the viscosity solution is given by* $u(x,y) = |x| \in H^1(\Omega)$.

We first approximate the example by partitioning Ω using an odd number of rectangles in both the x and y coordinate directions. Thus, the line $x = 0$ does



(a) $r = 1$. **(b)** $r = 2$.

Figure 4.3: Computed solutions for Example 4.3 using $r = 1$ with $\alpha = 10\,I$, $\gamma_{01} = \gamma_{03} = 100\,\mathbf{1}$, $\gamma_{02} = 200\,\mathbf{1}$, $\epsilon^i = -1\,\mathbf{1}$, $h =$ 2.357023e-01, and *fsolve* with initial guess $u_h^{(0)} = 0$ and $r = 2$ with $\alpha = 24\,I$, $\gamma_{01} = \gamma_{03} = 20\,\mathbf{1}$, $\gamma_{02} = 40\,\mathbf{1}$, $\epsilon^i = 0$, $h =$ 1.767767e-01 and 5 iterations of Algorithm 4.2 followed by *fsolve* with initial guess $u_h^{(0)} = 0$.

not correspond to an interior edge for any of the partitions. From Table 4.4 and Figure 4.4, we observe better than optimal rates of convergence in the L^2 norm for $r = 1$, using the fact that $u \in H^1(\Omega)$. Partitioning the domain into 64 uniform rectangles such that the line $x = 0$ always corresponds to an interior edge, we recover the exact solution for $r = 1$, as seen in Figure 4.5. For such a partition, we have $u \in V^h$.

Table 4.4: Rates of convergence for Example 4.4 using $\alpha = 100\,I$, $\gamma_{01} = \gamma_{03} = 20\,\mathbf{1}$, $\gamma_{02} = 40\,\mathbf{1}$, $\epsilon^i = 0$, and 10 iterations of Algorithm 4.2 followed by *fsolve* with initial guess $u_h^{(0)} = 0$. Note, the line $x = 0$ does not correspond to an interior edge for any of the partitions.

r	Norm	$h =$ 9.43e-01	$h =$ 5.66e-01		$h =$ 4.04e-01		$h =$ 2.57e-01	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	1.52e-01	7.25e-02	1.45	4.39e-02	1.49	2.24e-02	1.49
	L^∞	2.81e-01	1.73e-01	0.96	1.22e-01	1.02	7.73e-02	1.02

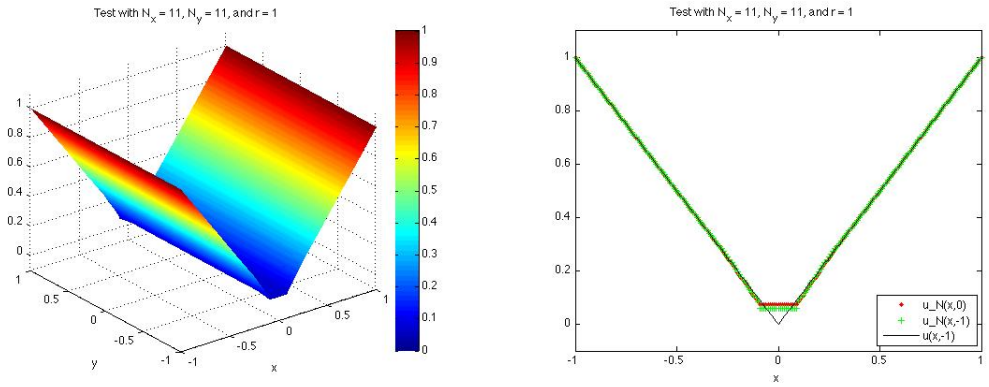


Figure 4.4: Computed solution for Example 4.4 using $r = 1$, $\alpha = 100\, I$, $\gamma_{01} = \gamma_{03} = 20\, \mathbf{1}$, $\gamma_{02} = 40\, \mathbf{1}$, $\epsilon^i = 0$, $h =$ 2.571297e-01 and 10 iterations of Algorithm 4.2 followed by *fsolve* with initial guess $u_h^{(0)} = 0$.

Example 4.5. *Consider the stationary Hamilton-Jacobi-Bellman problem*

$$\inf_{0\le\theta(x)\le 1}\left\{-\theta u_{xx} + \theta^2\, x^2\, u_x + \frac{1}{x}u + S(x)\right\} = 0, \qquad 1.2 < x < 4,$$
$$u(1.2) = 1.44 \ln 1.2, \quad u(4) = 16 \ln 4,$$

where

$$S(x) = \frac{4\ln(x)^2 + 12\ln(x) + 9 - 8x^4\ln(x)^2 - 4x^4\ln(x)}{4x^3\,[2\ln(x)+1]}.$$

This problem has the exact solution $u(x) = x^2 \ln x$, *which corresponds to* $\theta(x) = \frac{2\ln(x)+3}{2x^3[2\ln(x)+1]}$.

We solve this problem using various order elements and record the numerical results in Table 4.5 and Figure 4.6. Thus, our IPDG methods can also directly handle Hamilton-Jacobi-Bellman-type fully nonlinear PDEs as well.

Example 4.6. *Consider the stationary Hamilton-Jacobi-Bellman problem*

$$\min\{-\Delta u, -\Delta u/2\} = f \qquad \text{in } \Omega,$$
$$u = g \qquad \text{on } \partial\Omega,$$

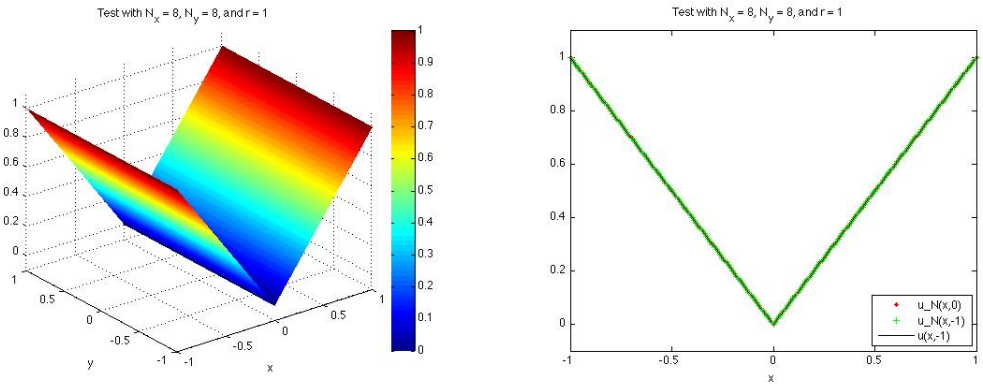


Figure 4.5: Computed solution for Example 4.4 using $r = 1$, $\alpha = 100\, I$, $\gamma_{01} = \gamma_{03} = 20\, \mathbf{1}$, $\gamma_{02} = 40\, \mathbf{1}$, $\epsilon^i = 0$, $h = 3.535534\text{e-}01$ and 10 iterations of Algorithm 4.2 followed by *fsolve* with initial guess $u_h^{(0)} = 0$. Note, the line $x = 0$ corresponds to only interior edges in the partition, and we have $u \in V^h$. The errors are given by $\|u_h - u\|_{L^\infty(\Omega)} = 6.88\text{e-}15$ and $\|u_h - u\|_{L^2(\Omega)} = 4.67\text{e-}15$. Thus, we capture the exact solution.

where $\Omega = (0, \pi) \times (-\pi/2, \pi/2)$,

$$f(x,y) = \begin{cases} 2\cos(x)\,\sin(y), & \text{if } (x,y) \in S, \\ \cos(x)\,\sin(y), & \text{otherwise,} \end{cases}$$

$S = (0, \pi/2] \times (-\pi/2, 0] \cup (\pi/2, \pi] \times (0, \pi/2)$, *and* g *is chosen such that the viscosity solution is given by* $u(x,y) = \cos(x)\,\sin(y)$.

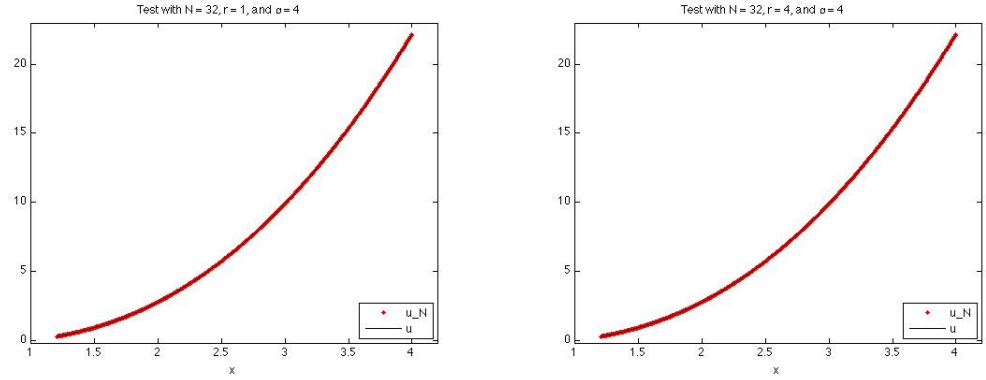
The results for approximating the problem with $r = 1$, $r = 2$, and $r = 3$ are recorded in Table 4.6 and Figure 4.7. Again, for $r = 1$, the calculated rates appear less than the predicted rates and, for $r = 2$, the calculated rates appear greater than the predicted rates. The calculated rages for $r = 3$ appear to agree with the predicted rate of 4 when averaged.

Table 4.5: Rates of convergence for Example 4.5 using $\alpha = 4$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $\overline{u}$.

r	Norm	$h = 2.8/4$	$h = 2.8/8$		$h = 2.8/16$		$h = 2.8/32$	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	3.5e-01	9.8e-02	1.83	2.6e-02	1.93	6.6e-03	1.97
	L^∞	3.9e-01	1.2e-01	1.70	3.4e-02	1.81	9.0e-03	1.91
2	L^2	9.1e-03	1.9e-03	2.28	4.2e-04	2.18	9.6e-05	2.11
	L^∞	9.9e-03	1.7e-03	2.53	3.6e-04	2.23	8.2e-05	2.15
3	L^2	3.5e-04	2.7e-05	3.69	1.9e-06	3.85	4.2e-07	2.14
	L^∞	5.1e-04	4.2e-05	3.61	3.3e-06	3.69	3.7e-07	3.15
4	L^2	2.5e-05	1.4e-06	4.14	7.7e-08	4.19	8.5e-09	3.18
	L^∞	3.3e-05	1.5e-06	4.46	7.6e-08	4.30	1.3e-08	2.51

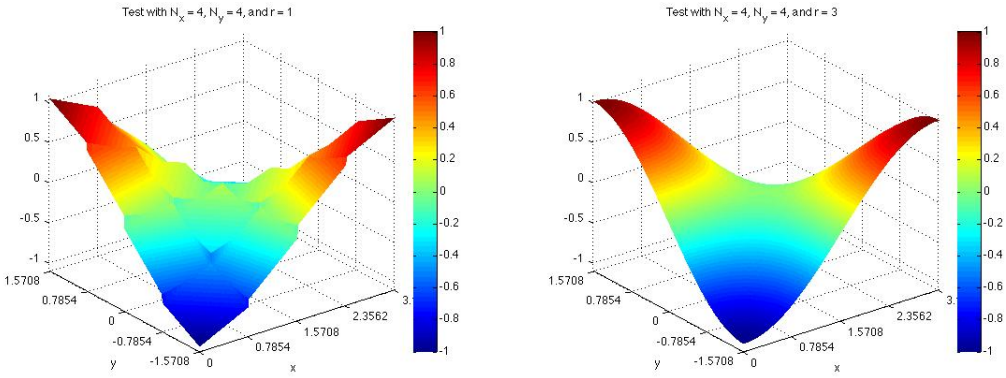
Table 4.6: Rates of convergence for Example 4.6 using $\alpha = 10\,\mathbf{1}$, $\gamma_{01} = \gamma_{03} = 20\,\mathbf{1}$, $\gamma_{02} = 40\,\mathbf{1}$, $\epsilon^i = 0$, and 4 iterations of Algorithm 4.2 followed by *fsolve* with initial guess $u_h^{(0)} = 0$.

r	Norm	$h = 2.22$e+00	$h = 1.48$e+00		$h = 1.11$e+00		$h = 8.89$e-01	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	4.86e-01	2.79e-01	1.37	1.92e-01	1.29	1.51e-01	1.09
	L^∞	2.86e-01	2.33e-01	0.51	1.69e-01	1.12	1.20e-01	1.55
2	L^2	1.97e-01	7.81e-02	2.28	3.83e-02	2.48	2.49e-02	1.92
	L^∞	1.51e-01	5.95e-02	2.29	3.02e-02	2.36	1.72e-02	2.52
3	L^2	7.60e-02	1.25e-02	4.46	7.77e-03	1.65	1.81e-03	6.53
	L^∞	5.56e-02	1.06e-02	4.08	5.74e-03	2.14	1.54e-03	5.89



(a) $r = 1$. **(b)** $r = 4$.

Figure 4.6: Computed solution for Example 4.5 using $\alpha = 4$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, $h =$ 1.250e-01, and *fsolve* with initial guess $\overline{u}$.



(a) $r = 1$. **(b)** $r = 3$.

Figure 4.7: Computed solution for Example 4.6 using $h = 1.11$, $\alpha = 10\,\mathbf{1}$, $\gamma_{01} = \gamma_{03} = 20\,\mathbf{1}$, $\gamma_{02} = 40\,\mathbf{1}$, $\epsilon^i = 0$, and 4 iterations of Algorithm 4.2 followed by *fsolve* with initial guess $u_h^{(0)} = 0$.

4.4.2 Parabolic Problems

We now implement the proposed fully discrete forward and backward Euler IPDG methods for approximating fully nonlinear parabolic equations of the form (4.13). While the above formulation makes no attempt to formally quantify a CFL condition for the forward Euler method, our test problems generally require $\Delta t = \mathcal{O}(h^2)$ to ensure the stability. In fact, the constant for the CFL condition appears to decrease as the order of the elements increases. Below we implement both the implicit and explicit methods for each test problem. However, we make no attempt to classify and compare the efficiency of the two methods. Instead, we focus on testing and demonstrating the usability of both fully discrete schemes and their promising potentials. For explicit tests, we record the parameter κ_t which serves as the scaling constant for the CFL condition enforced by the expression $\Delta t = \kappa_t h^2$. For implicit tests, we record computed solutions with various time steps Δt.

Example 4.7. *Consider the problem*

$$\begin{aligned} u_t - u_{xx}\, u &= f && \text{in } \Omega \times (0,1], \\ u &= g && \text{on } \partial\Omega \times (0,1], \\ u &= u_0 && \text{in } \Omega \times \{0\}, \end{aligned}$$

where $\Omega = (0,1)$, $f(x,t) = -\frac{1}{2}x^2 - t^4 + 4\,t^3 - 1$, *and* g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = 0.5\,x^2 + t^4 + 1$.

The given problem is actually quasi-linear, not fully nonlinear. The numerical results for the fully discrete forward Euler method are presented in Table 4.7 and Figure 4.8, and the results for the backward Euler method are shown in Table 4.8 and Figure 4.9. We observe that the errors for the backward Euler method are dominated by the relatively small size of the time step when compared to the forward Euler method. For smaller time step sizes, the errors are similar. However, the backward Euler method appears unstable for $\kappa_t > 0.01$.

Table 4.7: Rates of convergence in space for Example 4.7 at time $t = 1$ using the forward Euler method with $\kappa_t = 0.002$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and $u_h^0 = \mathcal{P}_h u_0$. The scheme appears unstable for $r = 2, 3$ when $\kappa_t = 0.01$.

r	Norm	$h = 1/4$	$h = 1/8$		$h = 1/16$		$h = 1/32$	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	5.7e-03	1.4e-03	1.98	3.7e-04	1.99	9.2e-05	1.99
	L^∞	7.9e-03	2.0e-03	1.99	5.0e-04	1.99	1.3e-04	1.99
2	L^2	3.3e-05	8.2e-06	2.00	2.1e-06	2.00	5.1e-07	2.00
	L^∞	4.5e-05	1.1e-05	2.00	2.8e-06	2.00	7.1e-07	2.00
3	L^2	3.3e-05	8.2e-06	2.00	2.1e-06	2.00	5.1e-07	2.00
	L^∞	4.5e-05	1.1e-05	2.00	2.8e-06	2.00	7.1e-07	2.00

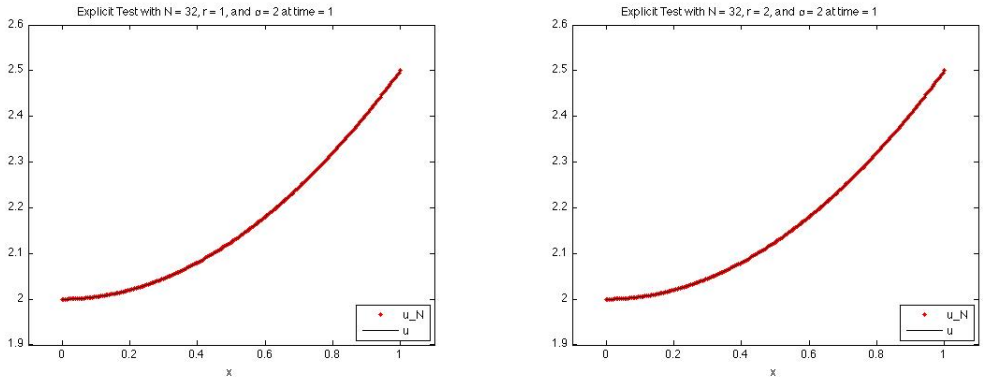
We now consider the error for the approximation resulting from using Euler time stepping methods. Note that the solution u is a quadratic in space. Letting $r = 2$, we limit the approximation error almost entirely to the time discretization scheme. In fact, setting $t = 0$ and solving the stationary form of the PDE, we have

$$\|u - u_h\|_{L^2((0,1))} \approx 1.6 \times 10^{-9} \quad \text{and} \quad \|u - u_h\|_{L^\infty((0,1))} \approx 2.4 \times 10^{-9}$$

using the elliptic solver with $h = 0.25$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 1$, $\gamma_{02} = 1.1$, and initial guess given by the secant line of the boundary data. Then, approximating the problem for varying Δt, we have the results recorded in Table 4.9 for the forward Euler method and in Table 4.10 for the backward Euler method. We observe that the convergence rate in time appears to have order 1 as expected.

Example 4.8. *Consider the problem*

$$\begin{aligned} u_t - u_x \ln\big(u_{xx} + 1\big) &= f && \text{in } \Omega \times (0, 3.1], \\ u &= g && \text{on } \partial\Omega \times (0, 3.1], \\ u &= u_0 && \text{in } \Omega \times \{0\}, \end{aligned}$$



(a) $r = 1$. **(b)** $r = 2$.

Figure 4.8: Computed solutions at time $t = 1$ for Example 4.7 using the forward Euler method with $\kappa_t = 0.002$, $h = 3.125\text{e-}02$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and $u_h^0 = \mathcal{P}_h u_0$.

where $\Omega = (0, 2)$, $f(x,t) = -e^{(t+1)x}\Big(x - (t+1)\ln\big((t+1)^2 e^{(t+1)x} + 1\big)\Big)$, *and* g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = e^{(t+1)x}$.

Notice, the exact solution u cannot be factored into the form $u(x,t) = G(t)\,Y(x)$ for some functions G and Y. The numerical results for the fully discrete forward Euler method are recorded in Table 4.11 and Figure 4.10, and the results for the backward Euler method are given in Table 4.12 and Figure 4.11. The error appears to be dominated by the low order time discretization given the relatively large value for Δt in the backward Euler test. However, when using a smaller value of Δt for the forward Euler test, we achieved a higher order of accuracy. We remark that even for $\Delta t = 0.005h^2$, the forward Euler scheme is not stable for $h = 0.25$ and $r = 1$.

Example 4.9. *Consider the problem*

$$\begin{aligned} u_t - \min_{\theta\in\{1,2\}} \Big\{ A_\theta\, u_{xx} - C\,\cos(t)\,\sin(x) - \sin(t)\,\sin(x) \Big\} = 0 \qquad & \text{in } \Omega \times (0, 3.1], \\ u = g \qquad & \text{on } \partial\Omega \times (0, 3.1], \\ u = u_0 \qquad & \text{in } \Omega \times \{0\}, \end{aligned}$$

Table 4.8: Rates of convergence in space for Example 4.7 at time $t = 1$ using the backward Euler method with $\Delta t = 0.001$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = 1/4$	$h = 1/8$		$h = 1/16$	
		Error	Error	Order	Error	Order
1	L^2	4.4e-03	9.6e-04	2.20	1.8e-04	2.40
	L^∞	9.4e-03	2.4e-03	2.00	5.9e-04	2.00
2	L^2	2.6e-04	2.6e-04	-0.00	2.6e-04	-0.00
	L^∞	3.6e-04	3.6e-04	-0.00	3.6e-04	-0.00
3	L^2	2.6e-04	2.6e-04	-0.00	2.6e-04	-0.00
	L^∞	3.6e-04	3.6e-04	-0.00	3.6e-04	-0.00

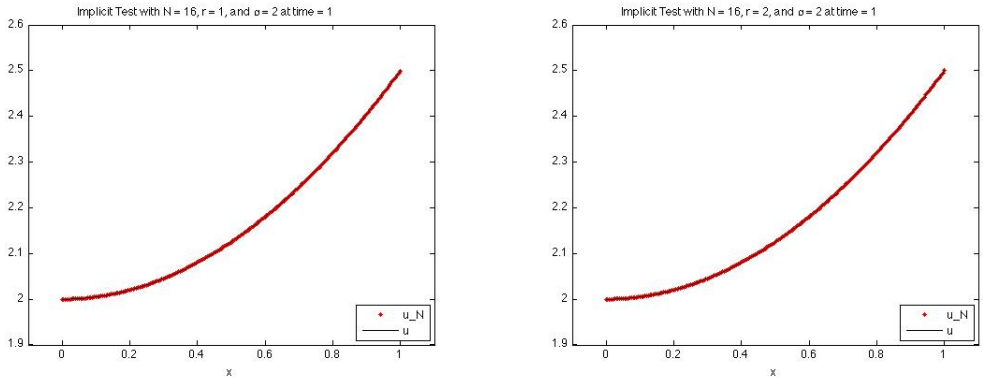
Table 4.9: Rates of convergence in time for Example 4.7 at time $t = 1$ using the forward Euler method with $h = 6.250\text{e-}02$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 1$, $\gamma_{02} = 1.1$, $\epsilon^i = 0$, and $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$\kappa_t = 0.008$	$\kappa_t = 0.004$		$\kappa_t = 0.002$		$\kappa_t = 0.001$	
		Error	Error	Order	Error	Order	Error	Order
2	L^2	8.2e-06	4.1e-06	1.00	2.1e-06	1.00	1.0e-06	1.00
	L^∞	1.1e-05	5.7e-06	1.00	2.8e-06	1.00	1.4e-06	1.00

where $\Omega = (0, 2\pi)$, $A_1 = 1, A_2 = \frac{1}{2}$,

$$C(x,t) = \begin{cases} 1, & \text{if } 0 < t \leq \frac{\pi}{2} \text{ and } 0 < x \leq \pi \text{ or } \frac{\pi}{2} < t \leq \pi \text{ and } \pi < x < 2\pi, \\ \frac{1}{2}, & \text{otherwise,} \end{cases}$$

and g *and* u_0 *are chosen such that the viscosity solution is given by* $u(x,t) = \cos(t)\sin(x)$.



(a) $r = 1$. **(b)** $r = 2$.

Figure 4.9: Computed solutions at time $t = 1$ for Example 4.7 using the backward Euler method with $\Delta t = 0.001$, $h = 6.250\text{e-}02$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

Table 4.10: Rates of convergence in time for Example 4.7 at time $t = 1$ using the backward Euler method with $h = 2.500\text{e-}01$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 1$, $\gamma_{02} = 1.1$, $\epsilon^i = 0$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$\Delta t = 1/10$	$\Delta t = 1/20$		$\Delta t = 1/40$		$\Delta t = 1/80$	
		Error	Error	Order	Error	Order	Error	Order
2	L^2	2.4e-02	1.3e-02	0.93	6.4e-03	0.96	3.2e-03	0.98
	L^∞	3.3e-02	1.7e-02	0.93	8.8e-03	0.96	4.5e-03	0.98

Notice that this problem involves an optimization over a finite dimensional set, and the viscosity solution corresponds to

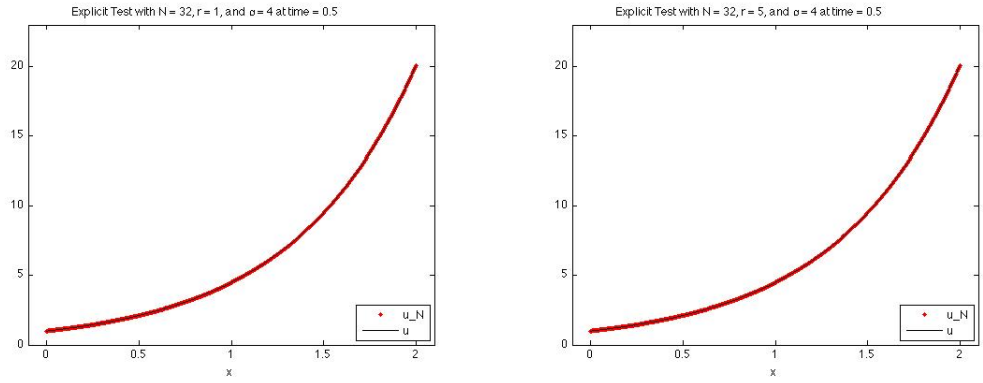
$$\theta(x,t) = \begin{cases} 1, & \text{if } c(x,t) = 1, \\ 2, & \text{if } c(x,t) = 2. \end{cases}$$

The numerical results are recorded in Table 4.13 and Figure 4.12 for the fully discrete forward Euler method and in Table 4.14 and Figure 4.13 for the fully discrete backward Euler method. We observe that the accuracy of the implicit method

Table 4.11: Rates of convergence in space for Example 4.8 at time $t = 3.1$ using the forward Euler method with $\kappa_t = 0.0025$, $\alpha = 4$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and $u_h^0 = \mathcal{P}_h u_0$. The scheme appears unstable for $\kappa_t = 0.005$.

r	Norm	$h = 1/2$	$h = 1/4$		$h = 1/8$		$h = 1/16$	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	5.0e-01	3.6e-02	1.73	1.2e-02	1.32	3.6e-03	1.67
	L^∞	8.2e-01	2.8e-01	1.57	1.0e-01	1.47	3.1e-02	1.69
2	L^2	4.5e-02	1.2e-02	1.89	3.3e-03	1.87	8.7e-04	1.93
	L^∞	6.0e-02	1.4e-02	2.11	3.6e-03	1.96	9.0e-04	1.98
3	L^2	1.5e-03	2.8e-04	2.39	7.1e-05	1.98	1.8e-05	1.98
	L^∞	2.7e-03	3.5e-04	2.97	7.6e-05	2.21	1.8e-05	2.05
4	L^2	1.2e-03	2.9e-04	2.06	7.2e-05	2.02	1.8e-05	2.01
	L^∞	1.3e-03	3.0e-04	2.13	7.3e-05	2.02	1.8e-05	2.01
5	L^2	1.2e-03	2.9e-04	2.00	7.2e-05	2.00	1.8e-05	2.00
	L^∞	1.2e-03	2.9e-04	2.00	7.3e-05	2.00	1.8e-05	2.00

appears to suffer from the lower order accuracy of the Euler method. For $h = \frac{\pi}{8}$, the explicit method requires $\Delta t \approx 3.1 \times 10^{-4}$, while the implicit method is computed with $\Delta t = 0.062$. When Δt increases, the explicit method demonstrates instability.

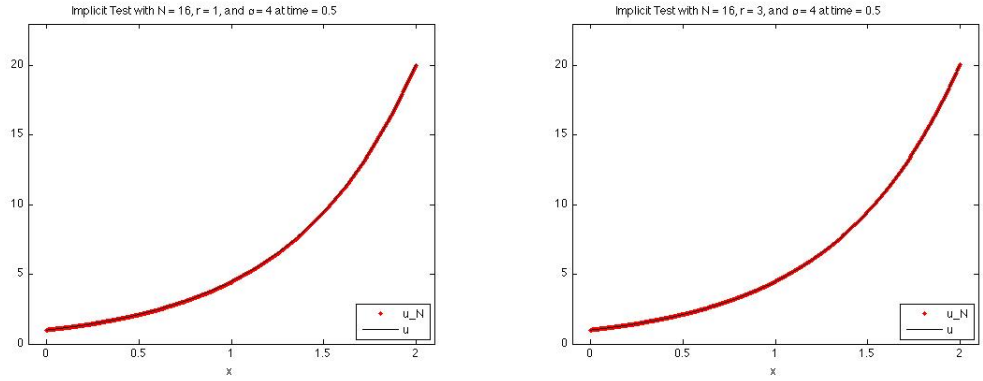


(a) $r = 1$. **(b)** $r = 5$.

Figure 4.10: Computed solutions at time $t = 3.1$ for Example 4.8 using the forward Euler method with $\kappa_t = 0.0025$, $h = 6.250\text{e-}02$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and $u_h^0 = \mathcal{P}_h u_0$.

Table 4.12: Rates of convergence in space for Example 4.8 at time $t = 3.1$ using the backward Euler method with $\Delta t = 0.0005$, $\alpha = 4$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = 1/2$	$h = 1/4$		$h = 1/8$	
		Error	Error	Order	Error	Order
1	L^2	4.2e-01	1.4e-01	1.54	4.6e-02	1.66
	L^∞	8.3e-01	2.4e-01	1.77	7.9e-02	1.63
2	L^2	7.3e-02	1.6e-02	2.21	3.0e-03	2.40
	L^∞	9.6e-02	1.8e-02	2.41	3.2e-03	2.49
3	L^2	2.8e-03	7.8e-04	1.82	9.1e-04	-0.22
	L^∞	5.6e-03	8.5e-04	2.71	9.2e-04	-0.11

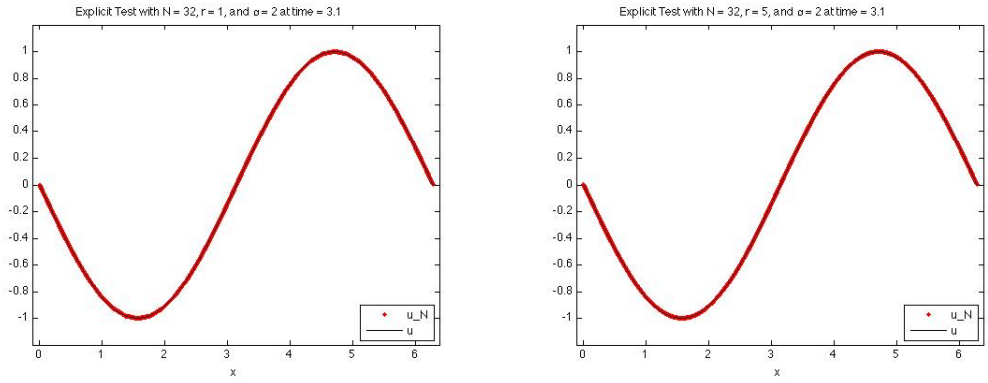


(a) $r = 1$. **(b)** $r = 3$.

Figure 4.11: Computed solutions at time $t = 1$ for Example 4.8 using the backward Euler method with $\Delta t = 0.0005$, $h = 1.250\text{e-}01$, $\alpha = 4$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

Table 4.13: Rates of convergence in space for Example 4.9 at time $t = 3.1$ using the forward Euler method with $\kappa_t = 0.002$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = \pi/2$	$h = \pi/4$		$h = \pi/8$		$h = \pi/16$	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	2.2e-01	5.3e-02	2.07	1.3e-02	2.02	3.3e-03	2.01
	L^∞	1.7e-01	4.8e-02	1.87	1.2e-02	1.98	3.1e-03	1.99
2	L^2	6.0e-02	1.6e-02	1.90	4.2e-03	1.94	1.1e-03	1.97
	L^∞	6.4e-02	1.5e-02	2.07	3.5e-03	2.13	8.2e-04	2.09
3	L^2	7.4e-03	6.9e-04	3.43	1.4e-04	2.32	3.5e-05	2.00
	L^∞	8.0e-03	5.6e-04	3.82	1.0e-04	2.46	2.3e-05	2.14
4	L^2	2.5e-03	5.7e-04	2.10	1.4e-04	2.03	3.5e-05	2.01
	L^∞	1.4e-03	3.5e-04	2.01	8.9e-05	1.98	2.2e-05	1.99
5	L^2	2.2e-03	5.6e-04	2.00	1.4e-04	2.00	3.5e-05	2.00
	L^∞	1.4e-03	3.6e-04	1.99	8.9e-05	2.00	2.2e-05	2.00

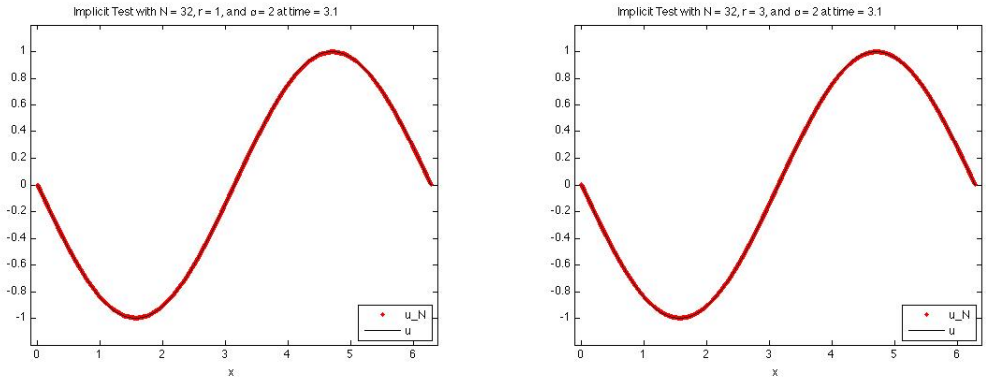


(a) $r = 1$. **(b)** $r = 5$.

Figure 4.12: Computed solutions at time $t = 3.1$ for Example 4.9 using the forward Euler method with $\kappa_t = 0.002$, $h = 1.963\text{e-}01$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and $u_h^0 = \mathcal{P}_h u_0$.

Table 4.14: Rates of convergence in space for Example 4.9 at time $t = 3.1$ using the backward Euler method with $\Delta t = 0.062$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

r	Norm	$h = \pi/2$	$h = \pi/4$		$h = \pi/8$		$h = \pi/16$	
		Error	Error	Order	Error	Order	Error	Order
1	L^2	1.7e-01	4.9e-02	1.82	1.4e-02	1.84	4.8e-03	1.50
	L^∞	1.5e-01	4.4e-02	1.78	1.3e-02	1.82	4.1e-03	1.60
2	L^2	8.0e-02	2.0e-02	2.00	5.9e-03	1.76	3.2e-03	0.87
	L^∞	7.0e-02	1.6e-02	2.14	4.0e-03	1.98	1.9e-03	1.06
3	L^2	1.1e-02	3.0e-03	1.91	2.8e-03	0.09	2.8e-03	0.00
	L^∞	8.1e-03	1.8e-03	2.16	1.8e-03	0.01	1.8e-03	0.00



(a) $r = 1$. **(b)** $r = 3$.

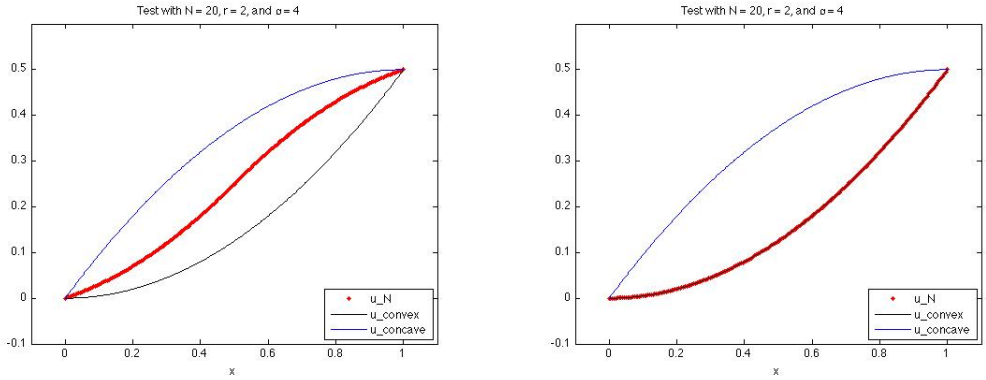
Figure 4.13: Computed solutions at time $t = 1$ for Example 4.9 using the backward Euler method with $\Delta t = 0.062$, $h = 1.963\text{e-}01$, $\alpha = 2$, $\gamma_{01} = \gamma_{03} = 2$, $\gamma_{02} = 2.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $u_h^0 = \mathcal{P}_h u_0$.

4.4.3 The Numerical Moment

We now perform a series of tests that are meant to explore the utility of the numerical moment term in the Lax-Friedrichs-like numerical operator defined by (4.4a). The numerical moment is the key to designing a consistent, g-monotone numerical operator. Thus, to better gauge the performance of the proposed IPDG methods, we must better understand the contribution of the numerical moment. Our main focus in this section will be on the Monge-Ampère equation, which is conditionally elliptic. In the following, unless otherwise stated, reported residual values are those recorded by *fsolve*, which corresponds to the ℓ^2 norm of the vector-valued system of equations evaluated at the current approximation value.

We begin by once again considering Example 4.2. We will show that the numerical moment provides a tool for addressing potential numerical artifacts, especially at the solver level. For the following tests, we let $\widehat{\mu}$ be defined by (3.44) and $\overline{u}$ denote the secant line formed by the boundary data in Example 4.2. Then, the test problem has two classical solutions, u^+ and u^-, infinitely many almost-everywhere C^1 artifacts such as $\widehat{\mu}$, and a unique viscosity solution u^+.

We first explore the possibility that the discretization contains numerical artifacts such as $\widehat{\mu}$. Let $r = 2$ in V^h. Then, $\widehat{\mu} \in C^1(\Omega) \cap V^h$. Thus, we have $[\widehat{\mu}] = [\widehat{\mu}_{x_k}] = 0$ on $\mathcal{E}_h^I$ for all $k = 1, 2, \ldots, d$ and $\widehat{\mu} = g$ on $\partial\Omega$. Therefore, whenever $x = 0.5$ is a node in the partition, we have $u_h = \widehat{\mu}$ and $P_{ih}(x) = 1$ if $x < 0.5$ and $P_{ih}(x) = -1$ if $x > 0.5$ for $i = 1, 2, 3$ is a numerical solution, and it follows that our discretization does have numerical artifacts when $r \geq 2$. We can see the presence of a numerical artifact in Figure 4.14. The function $\widehat{\mu}$ corresponds to a fixed point for the solver. However, if we slightly perturb the initial guess away from $\widehat{\mu}$, we see that Algorithm 4.2 converges to u^+. Unfortunately, the Newton algorithm *fsolve* does still converge to $\widehat{\mu}$ with the same slightly perturbed initial guess. Thus, for $r = 2$, our discretization does contain numerical artifacts that must be addressed at the solver level.



(a) $u_h^{(0)} = \widehat{\mu}$. **(b)** $u_h^{(0)} = 0.99\widehat{\mu} + 0.01\overline{u}$.

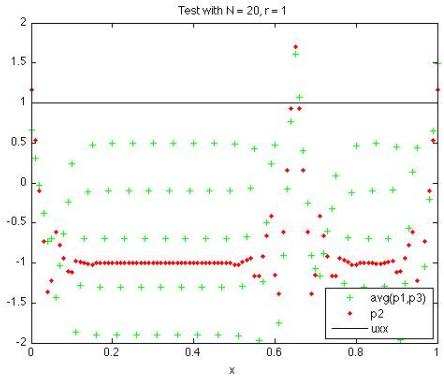
Figure 4.14: Computed solutions for Example 4.2 using $r = 2$, $\alpha = 4$, $h = 5.000\text{e-}02$, $\gamma_{01} = \gamma_{03} = 10.0$, $\gamma_{02} = 20.0$, $\epsilon^i = 0$, and Algorithm 4.2.

We now consider the presence of numerical artifacts when $r = 1$ in V^h. Then, $\widehat{\mu} \notin V^h$. Furthermore, the only C^1 function in V^h that satisfies the boundary condition is $\overline{u}$. Thus, we expect the numerical moment to have an effect since a good approximation for $u^{\pm}$ or $\widehat{\mu}$ must have jumps in the gradient along $\mathcal{E}_h^I$. Since we cannot determine analytically if there is a numerical solution in V^h that corresponds to $\widehat{\mu}$, we will explore the possibility numerically. To this end, we approximate Example 4.2

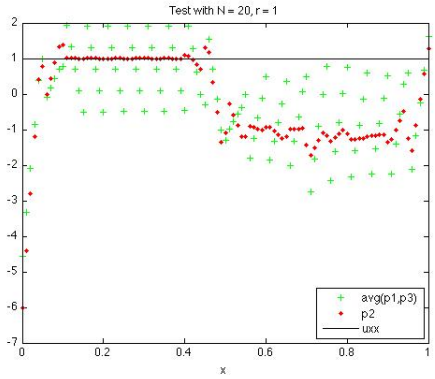
for $r = 1$ using *fsolve* and initial guesses that correspond to functions near $\widehat{\mu}$. The results can be found in Figure 4.15, where we plot the resulting values of P_{2h} and note that P_{2h} near -1 corresponds to u^- and P_{2h} near 1 corresponds to u^+. Observe, for the initial guess $u_h^{(0)} = \mathcal{P}_h\widehat{\mu}$, the solution appears to converge to a function near u^-. While the final approximation has a small residual, $\mathcal{O}(10^{-26})$, the last step for the solver was ineffective according to the error flags for *fsolve*. When approximating u^- with $r = 1$ by using a negative value for the coefficient of the numerical moment, the plot for P_{2h} is near -1 over the entire domain. Thus, while it is unclear if the approximation actually converged to u^-, it is clear that the approximation converged away from $\widehat{\mu}$. For the initial guess $u_h^{(0)} = \mathcal{P}_h\left(0.75\widehat{\mu} + 0.25\overline{u}\right)$, the solver does not find a root after 106 iterations. Instead, *fsolve* appears to be trapped in a small-residual well. After the 100th iteration, the residual is about 0.007. Thus, the discretization does not appear to have a numerical solution that corresponds to $\widehat{\mu}$ when $r = 1$. In contrast, when we set $\alpha = 0$, we clearly converge to a numerical artifact that corresponds to $\widehat{\mu}$. One last observation from Figure 4.15 is that $P_{1h} - 2P_{2h} + P_{3h}$ is nonzero in all three plots, as expected when using $r = 1$ paired with the lower regularity of $\widehat{\mu}$.

We now perform a series of three tests that deal specifically with the consequences of the numerical moment more at the solver level for Example 4.2. The first test will deal with the effect of various values of α paired with Newton solvers when the test problem has known numerical artifacts. The other two tests will be performed using $\gamma_{01} = \gamma_{02} = \gamma_{03}$. Then, we have $P_{2h} = \frac{P_{1h}+P_{3h}}{2}$, which in turn implies that the equation for P_{2h} is redundant in the formulation and the numerical moment should be zero upon convergence to a root.

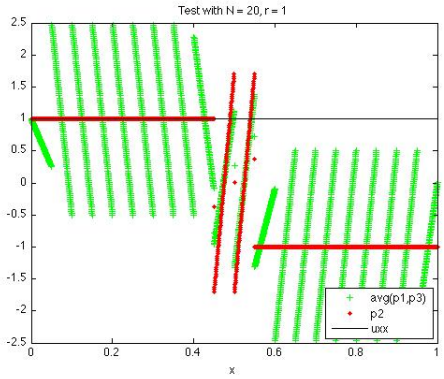
We first approximate Example 4.2 for $r = 2$ and various values of α using the full mixed formulation and *fsolve*. Observe, for $\alpha > 0$, our schemes should converge to u^+, and, for $\alpha < 0$, our schemes should converge to u^-, which is the unique viscosity solution of the PDE $u_{xx}^2 - 1 = 0$ with the same boundary data. However, for $\alpha = 0$, the scheme may converge to u^+, u^-, or a numerical artifact depending upon the initial



(a) $u_h^{(0)} = \mathcal{P}_h\widehat{\mu}$ and $\alpha = 4$.



(b) $u_h^{(0)} = \mathcal{P}_h\left(0.75\widehat{\mu} + 0.25\overline{u}\right)$ and $\alpha = 4$.



(c) $u_h^{(0)} = \mathcal{P}_h\widehat{\mu}$ and $\alpha = 0$.

Figure 4.15: Computed solutions for Example 4.2 using $r = 1$, $h =$ 5.000e-02, $\gamma_{01} = \gamma_{03} = 10.0$, $\gamma_{02} = 20.0$, $\epsilon^i = 0$, and *fsolve*.

guess used for the nonlinear solver. Note that while we cannot globally bound $F_{u_{xx}}$ for the operator $F(u_{xx}) = 1 - (u_{xx})^2$, we can locally bound $F_{u_{xx}}$. Thus, the necessary magnitude for α to allow for selective convergence depends on the initial guess and the solver. Without a global bound on $F_{u_{xx}}$, the numerical operator is only locally g-monotone (see Definition 4.2).

We let the initial guess be given by $u_h^{(0)} = \mathcal{P}_h\left(\frac{1}{3}\overline{u} + \frac{2}{3}\,u^-\right)$ and the initial guesses for P_{ih} be given by $P_{ih}^{(0)} = 0$ for $i = 1, 2, 3$. Thus, the initial guess is closer to u^-. From Table 4.15 and Figure 4.16, we see that the scheme converges to u^+ for $\alpha = 4$ and the scheme converges to u^- for $\alpha = 0$ and $\alpha = -4$ for the given parameters. If

we change the initial guess to $u_h^{(0)} = \mathcal{P}_h\left(\frac{1}{3}\overline{u} + \frac{2}{3}\,u^+\right)$, the scheme converges to u^+ for $\alpha = 0$ and $\alpha = 4$ and the scheme converges to u^- for $\alpha = -4$ for the given parameters. Furthermore, for $u_h^{(0)} = \mathcal{P}_h\overline{u}$, *fsolve* does not find a root for $\alpha = 0$, whereas the scheme converges to the desired solution for $\alpha = \pm 4$.

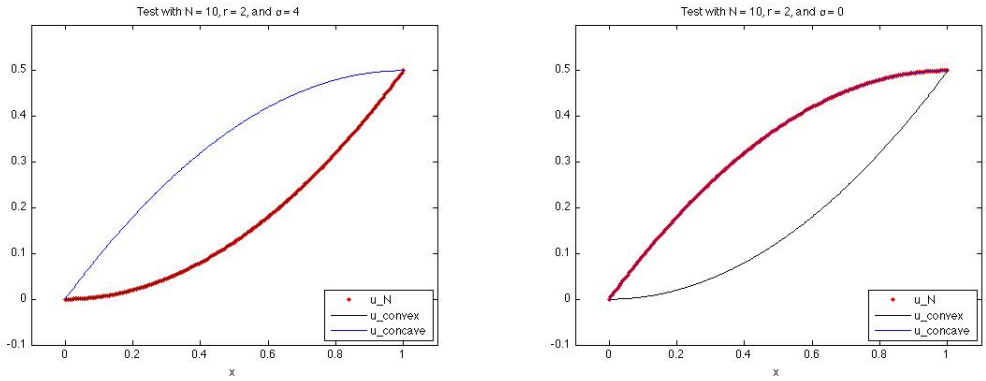
Table 4.15: Approximation errors for Example 4.2 using $r = 2$, $h =$ 1.000e-01, $\gamma_{01} = \gamma_{03} = 1.1$, $\gamma_{02} = 1.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $\mathcal{P}_h\left(\frac{1}{3}\overline{u} + \frac{2}{3}\,u^-\right)$.

Norm	$\alpha = 4$	$\alpha = 0$	$\alpha = -4$
L^2	2.5e-08	5.3e-10	3.7e-10
L^∞	3.3e-08	8.6e-10	5.7e-10

By the choice of α, we can enlarge the domain for which the numerical operator $\widehat{F}$ is increasing in P_{1h} and P_{3h} and decreasing in P_{2h}. Since the definition of ellipticity is based on the monotonicity of the operator, and the presence of numerical artifacts stems from whether the solution preserves the monotonicity of the operator, building monotonicity into the discretization is important when trying to preserve the nature of the operator we are approximating.

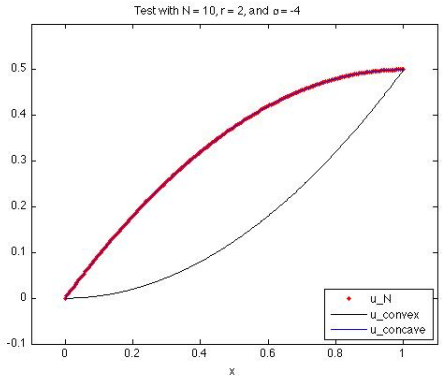
For the second test, we approximate Example 4.2 while plotting the norm of $P_{1h} - 2P_{2h} + P_{3h}$ after each iteration of *fsolve*. From Figure 4.17, we can see that even though we expect the numerical moment to be zero based upon the redundancy of the equation for P_{2h} given the equations for P_{1h} and P_{3h}, the Newton solver *fsolve* treats P_{1h}, P_{2h}, and P_{3h} as independent variables when searching for a root. The monotonicity of each variable appears to aid *fsolve* in the search for a root.

For the third test, we repeat the second test using Algorithm 4.2 instead of *fsolve* for the full mixed formulation. Let the initial guesses be given by $u_h = \mathcal{P}_h u^-$ and $P_{1h} = P_{2h} = P_{3h} = -0.99$. For $P_{2h} = -0.99$, F is increasing with respect to the Hessian argument while $\widehat{F}$ is decreasing for $\alpha > 0.99$. Since $F\,(-0.99) > 0$ and $\widehat{F}$ is decreasing for $P_{2h} \geq -1$ when $\alpha > 1$, we expect the splitting algorithm will



(a) $\alpha = 4$.

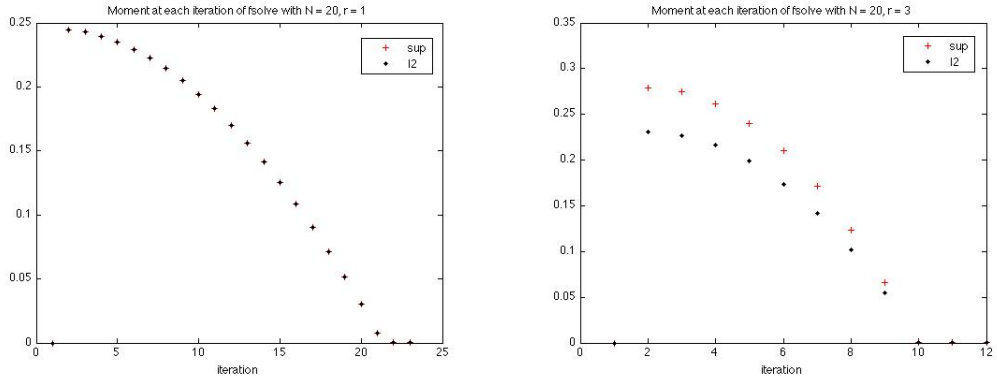
(b) $\alpha = 0$.



(c) $\alpha = -4$.

Figure 4.16: Computed solutions for Example 4.2 using $r = 2$, $h =$ 1.000e-01, $\gamma_{01} = \gamma_{03} = 1.1$, $\gamma_{02} = 1.5$, $\epsilon^i = 0$, and *fsolve* with initial guess $\mathcal{P}_h\left(\frac{1}{3}\overline{u} + \frac{2}{3}u^-\right)$.

move away from the concave root $P_{2h} = -1$. The numerical results are presented in Table 4.16 and Figure 4.18. We note that even with the initial guess close to u^-, the solver, with the aid of the numerical moment, converges to u^+. Similarly, the solver converges to u^+ when $P_{1h} = P_{2h} = P_{3h} > -1.0$ are used as initial guesses. For initial guesses $P_{1h} = P_{2h} = P_{3h} < -1.0$, the solver does not converge. In fact, we see the residuals measured by the L^∞ norm of $F(D_{2h}^2 u_h)$ diverging away from zero at an increasing rate. Thus, we see that even for the above simple solver, the monotonicity of $\widehat{F}$ provided by the numerical moment allows the scheme to either selectively converge to u^+ or diverge and find no solution. Hence, we again see the



(a) $r = 1$. **(b)** $r = 3$.

Figure 4.17: Iterative values for $P_{1h} - 2P_{2h} + P_{3h}$ when approximating Example 4.2 using $\alpha = 4$, $h =$ 5.000e-02, $\gamma_{01} = \gamma_{02} = \gamma_{03} = 2$, $\epsilon^i = 0$, and *fsolve* with initial guess $\overline{u}$.

benefit of including the numerical moment when tackling the presence of numerical artifacts when approximating viscosity solutions.

Next we consider the Monge-Ampère equation in two-dimensions. More specifically, we will once again consider Examples 4.3 and 4.4. We first consider Example 4.3. When approximating the given problem using a negative moment, our scheme still appears to converge to u, as seen in Figure 4.19. Thus, the negative moment does not appear to steer the solution towards the viscosity solution of the PDE $\det D^2u = f$, as it did using the LDG methods proposed in Chapter 3. However, we do note that when varying the initial guess, our scheme does not converge to a numerical artifact when using $r = 1$, which is not the case for standard discretizations that do not have a numerical moment (cf. [27]).

We also observe that the numerical moment plays a key role in preventing a Newton solver from encountering a singularity when solving the resulting system of nonlinear equations. In fact, for this example, *fsolve* does not converge when the numerical moment is not present, even with a good initial guess. We let $r = 1$, $\gamma_{02} = 100\,\mathbf{1}$, $\epsilon^2 = 0$, and $h =$ 2.828427e-01. Using the mixed formulation for only

Table 4.16: Rates of convergence for Example 4.2 using various values for α, $r = 1$, $\gamma_{01} = \gamma_{02} = \gamma_{03} = 2$, $\epsilon^i = 0$, and Algorithm 4.2 with initial guesses $u_h = u^-$ and $P_{1h} = P_{2h} = P_{3h} = -0.99$. For $\alpha \geq 1.1$, the scheme converges to u^+. For $\alpha \leq 0.99$, the scheme converges to u^-. When $\alpha = 1.0$, the scheme converges to u^+ for $h \geq \frac{1}{10}$ and the scheme converges to u^- for $h \leq \frac{1}{20}$.

α	Norm	$h = 1/5$	$h = 1/10$		$h = 1/20$		$h = 1/40$	
		Error	Error	Order	Error	Order	Error	Order
4	L^2	6.8e-03	1.7e-03	2.00	4.3e-04	2.00	1.0e-04	2.08
	L^∞	1.0e-02	2.5e-03	2.00	6.2e-04	2.00	1.6e-04	2.00
2	L^2	6.8e-03	1.7e-03	2.00	4.3e-04	2.01	9.8e-05	2.12
	L^∞	1.0e-02	2.5e-03	2.00	6.2e-04	2.00	1.6e-04	2.00
1.1	L^2	6.8e-03	1.7e-03	2.00	4.2e-04	2.01	9.7e-05	2.13
	L^∞	1.0e-02	2.5e-03	2.00	6.2e-04	2.00	1.6e-04	2.00
1	L^2	6.8e-03	1.7e-03	2.00	5.7e-04	1.58	8.2e-04	-0.53
	L^∞	1.0e-02	2.5e-03	2.00	9.4e-04	1.42	1.2e-03	-0.32
0.99	L^2	6.0e-03	9.7e-04	2.62	5.7e-04	0.77	8.2e-04	-0.53
	L^∞	9.9e-03	2.5e-03	2.00	9.4e-04	1.40	1.2e-03	-0.32
0	L^2	6.8e-03	1.7e-03	2.00	4.3e-04	1.99	1.1e-04	1.96
	L^∞	1.0e-02	2.5e-03	2.00	6.3e-04	1.99	1.6e-04	1.96

u_h and P_{2h} with $\alpha = 0$ and solving the resulting system of equations directly with *fsolve* has an initial residual of 357,315 with a residual of 197.73 after 50 iterations when the initial guess is given by $u_h^{(0)} = 0$ and has an initial residual of 90.35 with a residual of 1.41 after 50 iterations when the initial guess is given by $u_h^{(0)} = \mathcal{P}_h u$, the L^2 projection of the exact solution. Both attempts report the system of equations is close to singular, an error message that was not reported when performing the same tests with $\alpha \neq 0$.

We now consider Example 4.4, which features a solution in $H^1(\Omega) \setminus C^1(\Omega)$. Thus, the example does not fulfill the C^1 regularity assumption that was used in formulating the IPDG methods. We will perform a series of three tests, where we focus on both the choice of solver and the presence of a numerical moment. We will see that for this particular problem, the choice of the nonlinear solver has a larger impact

on whether or not the proposed IPDG methods successfully approximate the given viscosity solution.

We first approximate Example 4.4 using Algorithm 4.2 to solve the system of nonlinear equations. We can see in Figure 4.20 that the residuals measured by the L^∞ norm of $F[u_h]$ converge to zero quickly for $r = 1$ and appear to be converging towards zero slowly for $r = 2$. In fact, after 24 iterations, we have $\|u_h - u\|_{L^\infty(\mathcal{T}_h)} \approx$ 2.20e-04 and $\|u_h - u\|_{L^2(\mathcal{T}_h)} \approx$ 2.04e-04 for $r = 1$. After 1 iteration of Algorithm 4.2 with an initial guess of $u_h^{(0)} = 0$, we have $\left\| F\left(D_{2h}^2 u_h^{(1)}\right)\right\|_{L^\infty(\mathcal{T}_h)} \approx 13.54$. Furthermore, if $c_{k,\ell}^i$ denotes the coefficients for $\left[D_{ih}^2 u_h^{(1)}\right]_{k,\ell}$, $i = 1, 2, 3$, then we have $\left\|c_{1,1}^2 + c_{2,2}^2\right\|_{\ell^2} = 0$, as expected from the initial guess, and $\left\|c_{1,1}^1 + c_{1,1}^3 + c_{2,2}^1 + c_{2,2}^3\right\|_{\ell^2} \approx 22.9912$, indicating a nonzero numerical moment. From Figure 4.21, we can see that as Algorithm 4.2 iterates, the approximation does in fact become less smooth and appears to be converging towards the viscosity solution u.

We now approximate Example 4.4 using *fsolve* to solve the system of nonlinear equations. The results for using *fsolve* directly or *fsolve* after 20 iterations of Algorithm 4.2 can be found in Figure 4.22. We see that neither approximation converges to the viscosity solution, yet the residuals for *fsolve* are given by 3.34007e-26 after 11 iterations when we use *fsolve* directly and 1.63896e-26 after a maximum of 20 iterations when we first use Algorithm 4.2 to precondition the initial guess. Thus, using a Newton solver appears to yield C^0 numerical artifacts for the given problem. We do note that even with the small residuals, *fsolve* does return an error flag that indicates a possible lack of convergence for both tests. Also, while the first test that used *fsolve* fulfilled stopping criteria, the second test that used Algorithm 4.2 to precondition the initial guess had a trust-region radius for *fsolve* that was less than 5.0e-10 for the last 8 iterations causing the solver to stop prematurely.

We finally approximate Example 4.4 without using a numerical moment, as seen in Figure 4.23. When we do not have a numerical moment, *fsolve* does not converge after 25 iterations and has a residual of 103.035 with a residual of 109,660 corresponding

to the initial guess. We also note that *fsolve* reports the system of equations is singular or badly scaled after the first iteration and has a residual of 28,538 after the second iteration when not using a numerical moment. When using the numerical moment, our initial guess for *fsolve* has a residual of 37,158 and after 10 iterations converges with a residual of 1.33044e-26. Thus, we see that, for this example, using a numerical moment and preconditioning the initial guess for a Newton solver by first using Algorithm 4.2, we were able to approximate a degenerate problem that appears singular when using a straightforward discretization with a Newton solver.

From the above tests, we see that the numerical moment plays two major roles: it allows the scheme to converge for a wider range of initial guesses, especially when paired with the proper solver, and it enables the scheme to address the presence of numerical artifacts that can occur when approximating viscosity solutions. Given the form of the numerical moment, $\alpha : (P_{1h} - 2\,P_{2h} + P_{3h})$, these benefits are even more substantial given the way in which P_{1h}, P_{2h}, and P_{3h} are formed. The three variables only differ in their jump terms, and the entire numerical moment can be hard-coded using the jump-only representation derived in Section 4.1.3. When $\gamma_{01} = \gamma_{02} = \gamma_{03}$, the three different choices for the numerical fluxes (or jump terms) are all equivalent at the PDE level, and often the various jump formulations are presented as interchangeable when discretizing linear and quasi-linear PDEs using the IPDG methodology. Yet, for our schemes for fully nonlinear PDEs, we see that the three different choices of the numerical fluxes all play an essential role at the numerical level when combined to form the numerical moment, even in the degenerate case when $\gamma_{01} = \gamma_{02} = \gamma_{03}$.

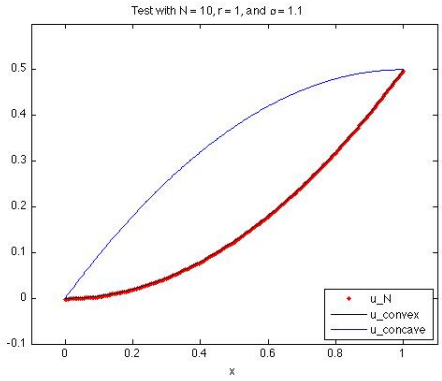
We end this section with a couple of remarks:

Remark 4.3.

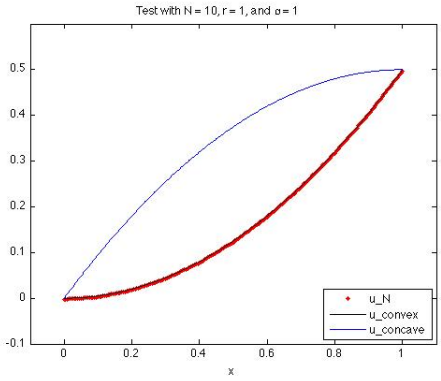
(a) The discretization techniques for fully nonlinear PDEs and the choice of solver for the resulting nonlinear systems of equations should not be considered entirely independent. We see in many tests that the addition of a numerical moment

yields a system of equations that is better suited for generic Newton solvers. However, the tests in Section 4.4.3 further indicate that the numerical moment has a much greater impact for approximating fully nonlinear PDEs when used in concert with an appropriate solver.

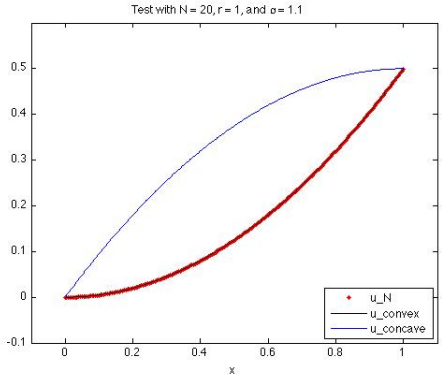
(b) *We see from the above tests for Example 4.4 that the numerical moment has potential to serve as an indicator function for adaptivity due to the fact it appears largest in areas where the viscosity solution is not regular.*



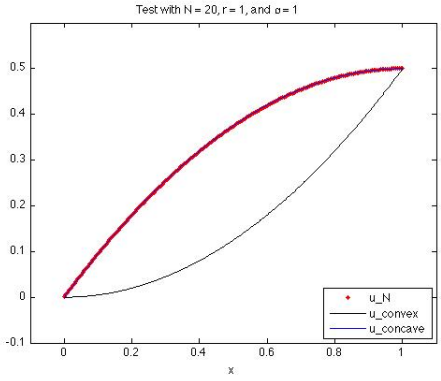
(a) $\alpha = 1.1$ and h=0.1.



(b) $\alpha = 1$ and h=0.1.



(c) $\alpha = 1.1$ and h=0.05.



(d) $\alpha = 1$ and h=0.05.

Figure 4.18: Computed solutions for Example 4.2 using various values for α, $r = 1$, $\gamma_{01} = \gamma_{02} = \gamma_{03} = 2$, $\epsilon^i = 0$, and Algorithm 4.2 with initial guesses $u_h = u^-$ and $P_{1h} = P_{2h} = P_{3h} = -0.99$. For $\alpha \geq 1.1$, the scheme converges to u^+. For $\alpha \leq 0.99$, the scheme converges to u^-. When $\alpha = 1.0$, the scheme converges to u^+ for $h \geq \frac{1}{10}$ and the scheme converges to u^- for $h \leq \frac{1}{20}$.

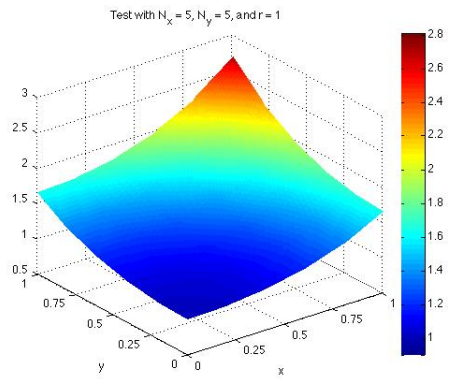


Figure 4.19: Computed solution for Example 4.3 using $r = 1$, $\alpha = -40\,I$, $\gamma_{01} = \gamma_{03} = 20\,\mathbf{1}$, $\gamma_{02} = 40\,\mathbf{1}$, $\epsilon^i = 0$, $h =$ 2.828427e-01 and 15 iterations of Algorithm 4.2 with initial guess $u_h^{(0)} = 0$.

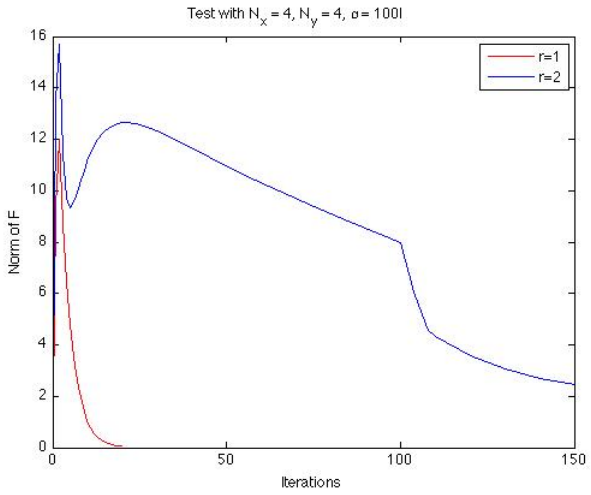
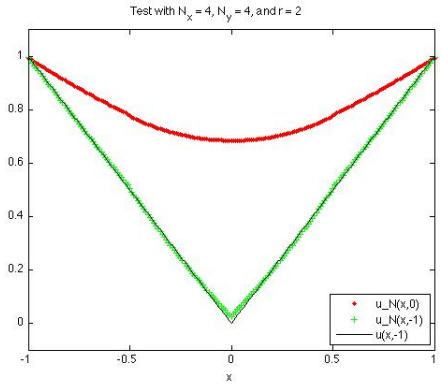
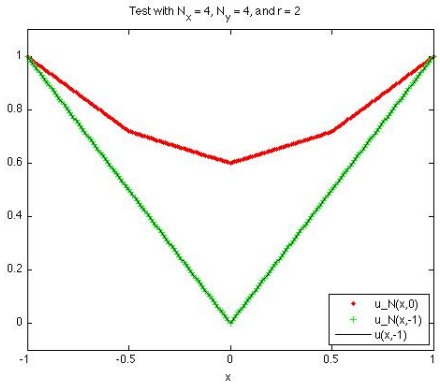


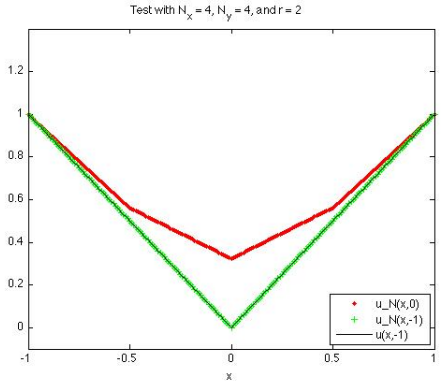
Figure 4.20: Computed residuals using the L^∞ norm of $F[u_h]$ for Example 4.4 using $\alpha = 100\,I$, $\gamma_{01} = \gamma_{03} = 20\,\mathbf{1}$, $\gamma_{02} = 40\,\mathbf{1}$, $\epsilon^i = 0$, $h =$ 7.071068e-01, and Algorithm 4.2 with initial guess $u_h^{(0)} = 0$.



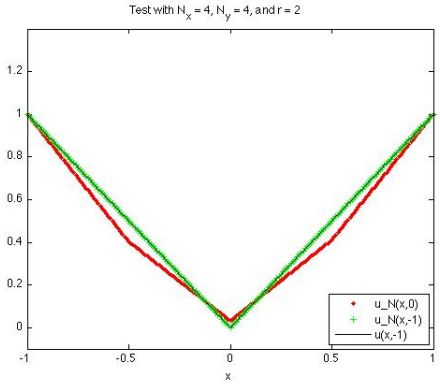
(a) u_h after 1 iteration.



(b) u_h after 20 iterations.



(c) u_h after 100 iterations.



(d) u_h after 150 iterations.

Figure 4.21: Computed solutions for Example 4.4 using $r = 2$, $\alpha = 100\,I$, $\gamma_{01} = \gamma_{03} = 20\,\mathbf{1}$, $\gamma_{02} = 40\,\mathbf{1}$, $\epsilon^i = 0$, $h = 7.071068\text{e-}01$, and Algorithm 4.2 with initial guess $u_h^{(0)} = 0$.

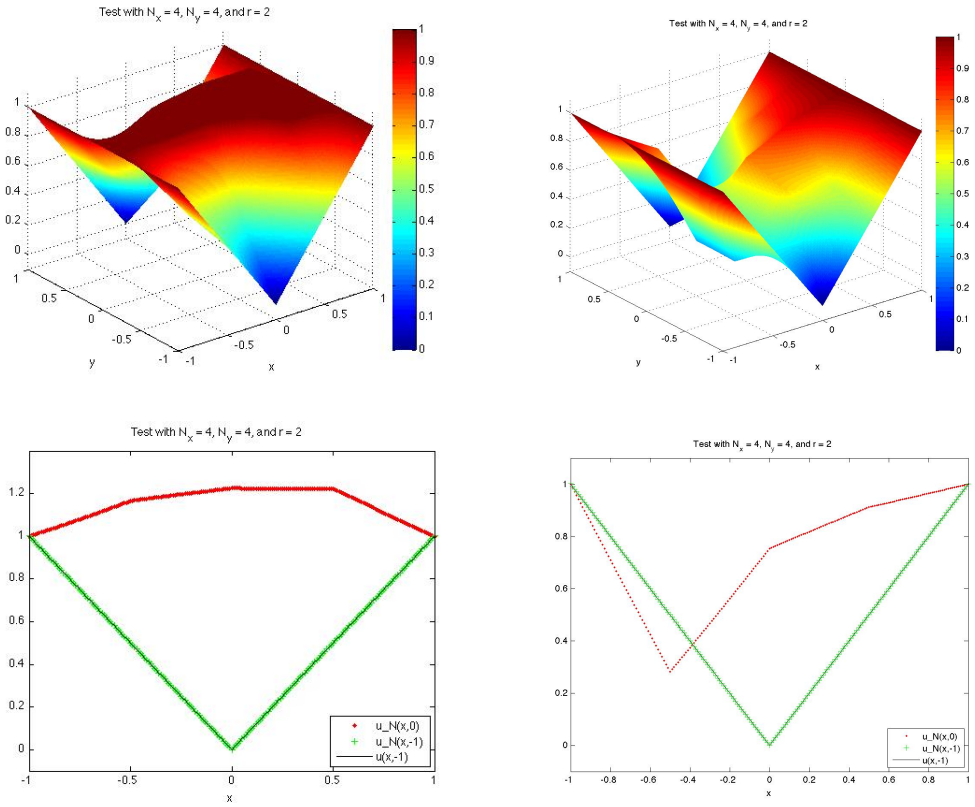
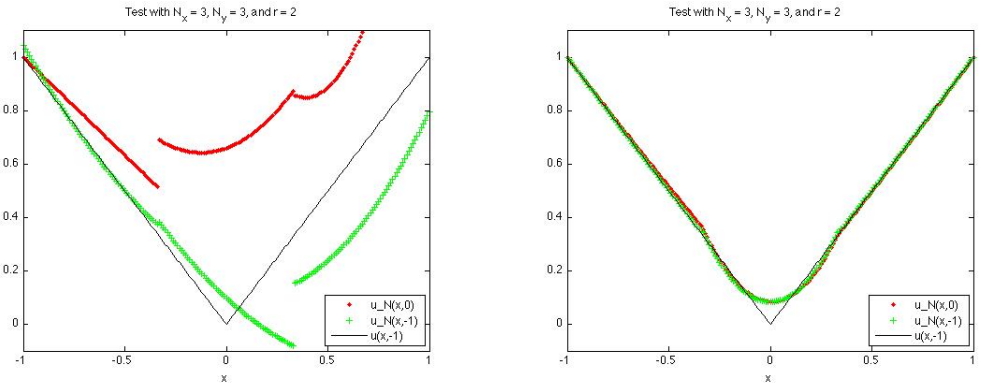


Figure 4.22: Computed solutions for Example 4.4 using $r = 2$, $\alpha = 100\, I$, $\gamma_{01} = \gamma_{03} = 20\, \mathbf{1}$, $\gamma_{02} = 40\, \mathbf{1}$, $\epsilon^i = 0$, $h = 7.071068\text{e-}01$, and *fsolve*. The left plots correspond to u_h with an initial guess $u_h^{(0)} = 0$, and the right plots correspond to u_h with the initial guess for *fsolve* given by the approximation after 20 iterations of Algorithm 4.2 with initial guess $u_h^{(0)} = 0$.



(a) $\alpha = 0$ **(b)** $\alpha = 100\, I$

Figure 4.23: Computed solutions for Example 4.4 using $r = 2$, $\gamma_{01} = \gamma_{03} = 20\,\mathbf{1}$, $\gamma_{02} = 40\,\mathbf{1}$, $\epsilon^i = 0$, $h = 9.428090\text{e-}01$, and initial guess $u_h^{(0)} = 0$. The left plot uses *fsolve* on the mixed system for only u_h and P_{2h}, and the right plot uses *fsolve* after 15 iterations of Algorithm 4.2.

Chapter 5

The Vanishing Moment Method for Hamilton-Jacobi-Bellman Equations

In this chapter, we propose an indirect method for approximating solutions of Hamilton-Jacobi-Bellman (HJB) equations. The previous chapters have been targeted towards directly approximating general fully nonlinear second order PDEs, and the methods developed were then tested on various problems including one- and two-dimensional HJB problems. However, in this chapter we consider another type of approximation methodology that explores the special structure of HJB equations. To this end, we will rewrite the HJB equation in a mixed form, and we will use the vanishing moment method (Section 1.3.1) to handle some of the numerical difficulties that will be associated with the resulting second order PDE. The proposed methodology will then be tested numerically in Section 5.3.

We now recall the HJB problem introduced in Section 1.4.2. Let $\Omega \subset \mathbb{R}^d$ be an open, bounded, convex domain. Suppose $T \in \mathbb{R}$ is positive and $\Theta \subset \mathbb{R}^m$. Then, we

consider the following HJB problem with Dirichlet boundary data:

$$u_t - \inf_{\theta(x,t)\in\Theta}\big(L_\theta u - f_\theta\big) = 0 \qquad \text{in } \Omega_T := \Omega \times (0,T], \tag{5.1a}$$

$$u(x,t) = g(x) \qquad \text{on } \partial\Omega_T := \partial\Omega \times (0,T], \tag{5.1b}$$

$$u(x,0) = u_0(x) \qquad \text{in } \Omega, \tag{5.1c}$$

where

$$L_\theta u := A^\theta : D^2 u + b^\theta \cdot \nabla u + c^\theta\, u \tag{5.2}$$

and $A^\theta : \Omega_T \to \mathbb{R}^{d\times d}$, $b^\theta : \Omega_T \to \mathbb{R}^d$, $c^\theta, f_\theta : \Omega_T \to \mathbb{R}$ for all $\theta \in \Theta$. Furthermore, we assume there exists $c > 0$ such that $\xi \cdot A^\theta(x,t)\xi \geq c|\xi|^2\ \forall \xi \in \mathbb{R}^d$ and $c^\theta(x,t) \geq 0$ for all $(x,t) \in \Omega_T$, $\theta \in \Theta$.

5.1 A Splitting Algorithm for the HJB Equation

We now consider developing an algorithm for solving problem (5.1). As in the previous chapters, we will first consider the time-independent problem. Then, to approximate (5.1), we again propose using the method of lines for the time discretization. Thus, for the remainder of the chapter, we consider the stationary HJB problem

$$-\inf_{\theta(x)\in\Theta}\big(L_\theta u - f_\theta\big) = 0 \qquad \text{in } \Omega, \tag{5.3a}$$

$$u = g \qquad \text{on } \partial\Omega, \tag{5.3b}$$

where L_θ is defined by (5.2) with t-independent coefficients.

Suppose problem (5.3) has a viscosity solution u. Then, we formally have (5.3a) is equivalent to the system of equations

$$
\begin{aligned}
-L_{\theta^*}u + f_{\theta^*} &= 0 && \text{in } \Omega, && \text{(5.4a)}\\
u &= g && \text{on } \partial\Omega, && \text{(5.4b)}\\
\theta^* &= \operatorname*{argmin}_{\theta\in\Theta}\left(L_\theta u - f_\theta\right) && \text{in } \Omega. && \text{(5.4c)}
\end{aligned}
$$

Therefore, a natural iterative algorithm for approximating the stationary HJB problem, (5.3), is given by the following algorithm:

Algorithm 5.1.

1. *Choose $\theta^{(0)}$, an initial guess for θ^*.*

2. *Successively solve the second order linear elliptic boundary value problem*

$$
\begin{aligned}
-L_{\theta^{(n)}}u^{(n)} + f_{\theta^{(n)}} &= 0 && \text{in } \Omega, && \text{(5.5a)}\\
u^{(n)} &= g && \text{on } \partial\Omega, && \text{(5.5b)}
\end{aligned}
$$

for $u^{(n)}$ and the optimization problem

$$
\theta^{(n+1)} = \operatorname*{argmin}_{\theta\in\Theta}\left(L_\theta u^{(n)} - f_\theta\right) \qquad \text{(5.6)}
$$

for $\theta^{(n+1)}$, for $n = 0, 1, 2, \ldots$.

Observe, Algorithm 5.1 has two major components, solving a second order linear elliptic boundary value problem in non-divergence form, (5.5), and an optimization problem, (5.6). For the remainder of the chapter we will focus on the former problem, approximating solutions to second order linear elliptic PDEs of non-divergence form, which is the case when A^θ is not differentiable. To this end, we define the linear

operator $L : H^2(\Omega) \to L^2(\Omega)$ by

$$Lv := -A : D^2 v = -\sum_{i,j=1}^{d} a_{i,j} v_{x_i x_j}, \tag{5.7}$$

and focus on the boundary value problem

$$Lu = f \qquad \text{in } \Omega, \tag{5.8a}$$

$$u = g \qquad \text{on } \partial\Omega \tag{5.8b}$$

with the following assumptions:

(i) Ω is open, connected, and bounded, with $\partial\Omega$ in $C^{1,1}$,

(ii) $A : \overline{\Omega} \to \mathbb{R}^{d\times d}$ with $A \in C^0\left(\overline{\Omega}\right)$,

(iii) $f : \Omega \to \mathbb{R}$ with $f \in L^2(\Omega)$,

(iv) $g : \partial\Omega \to \mathbb{R}$ with $g \in C^0(\partial\Omega)$,

(v) L is uniformly elliptic, i.e., there exists $\lambda, \Lambda > 0$ such that

$$\lambda |\xi|^2 \leq \xi \cdot A(x)\xi \leq \Lambda |\xi|^2 \qquad \forall\, \xi \in \mathbb{R}^d,\ x \in \Omega.$$

Then, by Chapter 9 of [32], there exists a unique function $u \in H^2(\Omega)$ such that u satisfies (5.8) almost everywhere in Ω. Thus, (5.8) has a unique *strong solution*. We note that we can make a weaker assumption in that $\partial\Omega$ only satisfy an exterior cone condition without consequence.

Remark 5.1.

(a) If we assume $A \in C^1(\overline{\Omega})$, then (5.8a) can be rewritten in divergence form using standard weak solution theory techniques. However, for $A \in C^0\left(\overline{\Omega}\right)$, we cannot rewrite (5.8a) in divergence form. We will see that problems of non-divergence form present many obstacles, even when the problem is linear.

(b) *If we assume* $A \in L^{\infty}(\Omega)$*, then there may not exist a strong solution to* (5.8)*. Instead, existence and uniqueness is only guaranteed under viscosity solution theory (cf. [32]).*

(c) *In order to apply the following techniques to HJB problems that arise from using the Bellman principle to recast stochastic optimal control problems, the following results in Section 5.2 need to be extended to problems of the type* (5.8) *with both lower-order terms and lower-regularity coefficient functions. Thus, the results in Section 5.2 must be extended in terms of viscosity solution theory before the methods can be used for a wider range of application problems.*

(d) *In the next section, we apply the vanishing moment method as a means to approximate the strong solution of* (5.8)*. Alternative methodologies for approximating second order linear elliptic PDE problems of non-divergence form can be found in [50] and the references therein. We note that the methods of Smears and Süli, while applicable for* $A \in L^{\infty}(\Omega)$*, make an alternative assumption that* A *satisfies a* Cordès condition*, i.e., there exists an* $\epsilon \in (0,1)$ *such that*

$$\frac{\sum_{i,j=1}^{d} a_{i,j}^2}{\left(\sum_{i=1}^{d} a_{i,i}\right)^2} \leq \frac{1}{d-1+\epsilon}$$

almost everywhere in Ω*. Such an assumption avoids having to deal with viscosity solution theory in its full generality.*

(e) *The finite difference methods of Chapter 2 have been shown to converge to the viscosity solution of* (5.8) *when* $A \in L^{\infty}(\Omega)$ *with* A *strictly diagonal. Thus, Algorithm 5.1 provides an alternative method for approximating HJB problems than was used in the numerical tests of Chapters 2,3, and 4.*

(f) *In fact, all of the direct methods developed in Chapters 2 - 4 are applicable to the linear problem* (5.8)*.*

5.2 The Vanishing Moment Method for Second Order Elliptic Problems of Non-Divergence Form

In this section, we apply the vanishing moment methodology (see Section 1.3.1) to the second order linear elliptic boundary value problem (5.8) of non-divergence form. To this end, we will develop a family of fourth order linear boundary value problems whose weak solutions converge to the strong solution of (5.8). Then, in order to approximate the solution of (5.8), we approximate the solutions of the fourth order problems that may be better suited for a wider range of numerical methodologies.

We now describe the family of fourth order linear boundary value problems. Pick $\epsilon > 0$. Define the linear operator $L^\epsilon : H^4(\Omega) \to L^2(\Omega)$ by

$$L^\epsilon v := \epsilon \Delta^2 v - A : D^2 v, \tag{5.9}$$

where Δ^2 denotes the biharmonic operator. Then, we consider the boundary value problem

$$L^\epsilon u^\epsilon = f \quad \text{in } \Omega, \tag{5.10a}$$

$$u^\epsilon = g \quad \text{on } \partial\Omega, \tag{5.10b}$$

$$\Delta u^\epsilon = 0 \quad \text{on } \partial\Omega. \tag{5.10c}$$

Note, (5.10) defines a family of fourth order linear boundary value problems parameterized by ϵ. Heuristically, we expect the contributions of the fourth order term $\epsilon \Delta^2 u^\epsilon$ and the high-order boundary condition $\Delta u^\epsilon = 0$ to "vanish" as $\epsilon \to 0$.

We now present a weak formulation of (5.10). Define $V_g \subset H^2(\Omega)$ by

$$V_g := \left\{ v \in H^2(\Omega) : v\big|_{\partial\Omega} = g \right\},$$

where V_g will denote the trial space for a weak formulation of (5.10). Note, if $g = 0$, then $V_0 = H^2(\Omega) \cap H_0^1(\Omega)$. Also, define the bilinear form $\mathcal{A} : V_0 \times V_0 \to \mathbb{R}$ by

$$\mathcal{A}(v, w) := \epsilon (\Delta v, \Delta w)_\Omega - \left(A : D^2 v, w\right)_\Omega, \tag{5.11}$$

where $(\cdot, \cdot)_\Omega$ denotes the L^2 inner product (see Section 5.2.1). Observe, there exists infinitely many $\tilde{u} \in C^4(\Omega) \cap C^0(\overline{\Omega})$ such that $\tilde{u}\big|_{\partial\Omega} = g$. Then,

$$\begin{aligned} L(u - \tilde{u}) &= f + L\tilde{u} \quad && \text{in } \Omega, \\ u - \tilde{u} &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Thus, for our analysis, we can assume $g = 0$ in (5.8b) without a loss of generality by associating u with $u - \tilde{u}$. Since u^ϵ is meant to approximate u, we also let $g = 0$ in (5.10b).

Suppose $u^\epsilon \in C^4(\Omega)$ is a classical solution to (5.10) and $w \in C^\infty$ such that $w\big|_{\partial\Omega} = 0$. Observe,

$$\begin{aligned} (L^\epsilon u^\epsilon, w)_\Omega &= \epsilon \left(\Delta^2 u^\epsilon, w\right)_\Omega - \left(A : D^2 u^\epsilon, w\right)_\Omega \\ &= \epsilon (\nabla \Delta u^\epsilon \cdot n, w)_{\partial\Omega} - \epsilon (\nabla \Delta u^\epsilon, \nabla w)_\Omega - \left(A : D^2 u^\epsilon, w\right)_\Omega \\ &= -\epsilon (\nabla \Delta u^\epsilon, \nabla w)_\Omega - \left(A : D^2 u^\epsilon, w\right)_\Omega \\ &= -\epsilon (\Delta u^\epsilon, \nabla w \cdot n)_{\partial\Omega} + \epsilon (\Delta u^\epsilon, \Delta w)_\Omega - \left(A : D^2 u^\epsilon, w\right)_\Omega \\ &= \epsilon (\Delta u^\epsilon, \Delta w)_\Omega - \left(A : D^2 u^\epsilon, w\right)_\Omega. \end{aligned}$$

Thus, we define a weak formulation for problem (5.10) as follows: Find $u^\epsilon \in V_0$ such that

$$\mathcal{A}(u^\epsilon, w) = (f, w)_\Omega \qquad \forall w \in V_0. \tag{5.12}$$

In the following sections we will show that (5.12) has a unique solution and that the weak solution converges to the strong solution of (5.8) in L^2 as $\epsilon \to 0$. We note that the existence and uniqueness results, as well as the convergence results in this section,

will be based upon Conjecture 5.1 found below and to be discussed in Section 6.3. In Section 5.3, we perform a series of numerical tests to support our results.

5.2.1 Notation

Before we continue, we first introduce some standard space notation. We will also introduce some special notation that will be convenient for the following analysis.

We first introduce notation for mollifier functions. Choose $\rho > 0$ and define Ω_ρ by

$$\Omega_\rho := \{x \in \Omega \mid \text{ dist}\,(x, \partial\Omega) > \rho\}\,.$$

Let η denote the standard mollifier, i.e.,

$$\eta\,(x) := \begin{cases} C\,\exp\left(\frac{1}{|x|^2-1}\right), & \text{if } |x| < 1, \\ 0, & \text{if } |x| \geq 1, \end{cases}$$

where the constant $C > 0$ is defined such that $\int_{\mathbb{R}^d} \eta\, dx = 1$. Define η_ρ by

$$\eta_\rho\,(x) := \frac{1}{\rho^d}\eta\left(\frac{x}{\rho}\right).$$

Then, $\eta_\rho \in C^\infty(\mathbb{R}^d)$ and $\int_{\mathbb{R}^d} \eta_\rho\, dx = 1$ with the support of η_ρ a proper subset of $B_\rho(0)$, the open ball of radius ρ centered at the origin.

If $f : \Omega \to \mathbb{R}$ is locally integrable, define its mollification by

$$f_\rho := \eta_\rho * f \qquad \text{in } \Omega_\rho,$$

that is,

$$f_\rho\,(x) = \int_\Omega \eta_\rho\,(x-y)\,f\,(y)\,dy = \int_{B_\rho(0)} \eta_\rho\,(y)\,f\,(x-y)\,dy$$

for all $x \in \Omega_\rho$. The following facts are standard, see [22]:

Theorem 5.1. *Let p be a nonnegative integer and $f : \Omega \to \mathbb{R}$ be locally integrable. Then, there holds*

(i) $f_\rho \in C^\infty(\Omega_\rho)$.

(ii) $f_\rho \to f$ *a.e. as* $\rho \to 0$.

(iii) if $f \in H^p_{loc}(\Omega)$, *then* $D^\alpha f_\rho = (D^\alpha f)_\rho$ *for all* $|\alpha| \leq p$.

(iv) if $f \in H^p_{loc}(\Omega)$, *then* $f_\rho \to f$ *in* $H^p_{loc}(\Omega)$.

(v) if $f \in H^p(\Omega)$, *then* $f_\rho \to f$ *in* $H^p(\Omega)$.

In the following we will have several generic constants. For consistency, we let C_P denote the optimal constant from Poincaré's inequality on Ω, C_I denote the optimal constant from the interpolation inequality on Ω, and C_α denote the optimal constant from the trace inequality on Ω, i.e. for $\alpha \in [0, 1/2]$,

$$\|v\|_{L^2(\partial\Omega)} \leq C_\alpha \|v\|^{1-\alpha}_{H^1(\Omega)} \|v\|^\alpha_{L^2(\Omega)}$$

for all $v \in H^1(\Omega)$, where α depends on the dimension d (see [1] for more details). Unless the dependency on α is shown explicitly, we use the weaker bound that corresponds to $\alpha = 0$. In this case, we denote the generic constant by C_T.

Let n be a positive integer such that $1 \leq n \leq d$. For matrices $B, C \in \mathbb{R}^{n\times n}$, define the matrix inner product $B : C$ by

$$B : C := \text{ tr } \left(BC^T\right) = \sum_{i,j=1}^{n} b_{ij} c_{ij},$$

which induces the Frobenius norm. If $B, C : \Omega \to \mathbb{R}^{n\times n}$, define the L^2 inner product $(B, C)_\Omega$ by

$$(B, C)_\Omega = \int_\Omega B : C \, dx.$$

Then, $\left(L^2\left(\Omega;\mathbb{R}^{n\times n}\right),(\cdot,\cdot)\right)$ forms a Hilbert space with the norm

$$\|B\|_{L^2(\Omega)} := \sqrt{(B,B)_\Omega}$$

for all $B \in L^2\left(\Omega;\mathbb{R}^{n\times n}\right)$.

We now develop some notation specific to problem (5.10). Let A be the coefficient matrix in the definition of the operator L. Since A is continuous on $\overline{\Omega}$ and $\overline{\Omega}$ is compact, A is uniformly continuous on $\overline{\Omega}$. Thus, for any $\delta_A > 0$, there exists $\rho > 0$ such that, if $|x-y| < \rho$, then $|A(x) - A(y)| < \delta_A$ for all $x, y \in \overline{\Omega}$.

Overlay $\mathbb{R}^d$ with a uniform grid of diameter $\rho/3$. Since $\overline{\Omega}$ is compact, there exists a finite number $M := M\left(\delta_A\right) > 0$ such that $\overline{\Omega} \subset \bigcup_{j=1}^{M} \overline{O}_j$, where O_j denotes an arbitrary open cell formed by the grid. Define $\Omega_j := O_j \cap \Omega$ for all $j = 1, 2, \ldots, M$. Then, $\overline{\Omega} = \bigcup_{j=1}^{M} \overline{\Omega}_j$ and $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$.

Choose x_j to be the barycentric center of Ω_j and define $A_j := A\left(x_j\right)$ for all $j = 1, 2, \ldots, M$. Define sets consisting of the interior and exterior surfaces of the partition by

$$\mathcal{E}^{\mathrm{I}} := \left\{ \Gamma_{i,j} = \partial\Omega_i \cap \partial\Omega_j \text{ for some } i, j \in \{1, 2, \ldots, M\} \right\}$$

and

$$\mathcal{E}^{\mathrm{B}} := \left\{ \Gamma_j = \partial\Omega \cap \partial\Omega_j \text{ for some } j \in \{1, 2, \ldots, M\} \right\}.$$

Then, we let $\overline{A} : \Omega \setminus \mathcal{E}^{\mathrm{I}} \to \mathbb{R}^{n\times n}$ be a piecewise constant matrix-valued function defined by

$$\overline{A}\left(x\right) := A_j \text{ for } x \in \Omega_j, \quad j = 1, 2, \ldots, M.$$

Lastly, we define $A^s := \frac{1}{2}\left(A + A^T\right)$ and $A^a := \frac{1}{2}\left(A - A^T\right)$. Then, A^s is symmetric, A^a is anti-symmetric, and $A = A^s + A^a$. Furthermore, A^s and A^a are uniformly continuous on $\overline{\Omega}$.

We end this section by stating two conjectures. The first conjecture will be used throughout the following sections where we address the issues of existence, uniqueness, and convergence for (5.10) in the semi-norm for H^1. The second conjecture will be used as part of a duality argument to derive rates of convergence in L^2.

Conjecture 5.1. *Let n_j denote the unit outward normal vector for Ω_j. Then, there exists a constant $C_A > 0$ independent of ρ such that*

$$\sum_{j=1}^{M} \left(A_j \nabla v \cdot n_j, v\right)_{\partial\Omega_j} \leq \delta_A \, C_A \, \|v\|^2_{H^2(\Omega)}$$

for all $v \in H^2(\Omega)$ and

$$\sum_{j=1}^{M} \left(A_j \nabla v \cdot n_j, \Delta v\right)_{\partial\Omega_j} \leq \delta_A \, C_A \, \|v\|^2_{H^3(\Omega)} \,,$$
$$\sum_{j=1}^{M} \left(A_j \nabla v, D^2 v \, n_j\right)_{\partial\Omega_j} \leq \delta_A \, C_A \, \|v\|^2_{H^3(\Omega)}$$

for all $v \in H^3(\Omega)$.

Conjecture 5.2. *Let n_j denote the unit outward normal vector for Ω_j. Then, there exists a constant $C_B > 0$ independent of ρ such that*

$$\sum_{j=1}^{M} \left(A_j \nabla v \cdot n_j, w\right)_{\partial\Omega_j} \leq \delta_A \, C_B \, \left(\|v\|^2_{H^2(\Omega)} + \|w\|^2_{H^2(\Omega)}\right)$$

for all $v, w \in H^2(\Omega)$.

The above conjectures will be used in a-priori estimates, and as such will not necessarily have to hold for all of $H^2(\Omega)$ and $H^3(\Omega)$. We also note that the proof of the conjectures may require the partition of Ω be comprised of sets that all have similar sizes and shapes. The above partitioning method may produce non-similar sets near the boundary of Ω, and thus need to be revised. A more detailed discussion of the conjectures can be found in Section 6.3.

5.2.2 Existence and Uniqueness

We now show that (5.10) has a unique solution using the abstract Galerkin method. Let $\{\phi_j\}_{j=1}^{\infty}$ denote the eigenfunctions of the Laplacian operator Δ, that is,

$$\Delta\phi_j = \mu_j\,\phi_j \qquad \text{in } \Omega, \tag{5.13a}$$

$$\phi_j = 0 \qquad \text{on } \partial\Omega, \tag{5.13b}$$

where $\{\mu_j\}_{j=1}^{\infty} \subset \mathbb{R}_+$ are the corresponding eigenvalues. It is well known that $\{\phi_j\}_{j=1}^{\infty}$ forms an orthonormal basis of $L^2(\Omega)$ with $\phi_j \in C^{\infty}(\Omega) \cap H_0^1(\Omega)$ for all $j \geq 1$, (cf. [22]).

Define

$$V_N := \text{span}\,\{\phi_1, \phi_2, \ldots, \phi_N\} \tag{5.14}$$

for $N \geq 1$. Then, $V_N \subset H_0^1(\Omega) \cap C^{\infty}(\Omega)$. We first show there exists a unique $u_N^{\epsilon} \in V_N$ such that

$$\mathcal{A}(u_N^{\epsilon}, v) = (f, v)_{\Omega} \tag{5.15}$$

for all $v \in V_N$, where $\mathcal{A}$ is defined by (5.11). To this end we will derive an a-priori estimate for solutions of (5.15). In the proof, we use the following fact found in [32]:

Theorem 5.2. *Let $v \in V_0$. Then*

$$\left\|D^2 v\right\|_{L^2(\Omega)} \leq C_0 \left\|\Delta v\right\|_{L^2(\Omega)}$$

and, if $v \in H^3(\Omega) \cap V_0$,

$$\left\|D^3 v\right\|_{L^2(\Omega)} \leq C_0 \left\|\nabla\Delta v\right\|_{L^2(\Omega)}$$

for some constant $C_0 > 0$.

We are now ready to prove the existence of solutions to (5.15) in V_N using the following a-priori estimate:

Theorem 5.3. *Assume u_N^ϵ is a solution of (5.15) in V_N. Then,*

$$\epsilon \left\| D^2 u_N^\epsilon \right\|_{L^2(\Omega)}^2 + \lambda\, c_1 \left\| \nabla u_N^\epsilon \right\|_{L^2(\Omega)}^2 \leq \frac{c_2}{\lambda} \left\| f \right\|_{L^2(\Omega)}^2, \tag{5.16}$$

where c_1, c_2 denote positive constants which are independent of N, f, λ, and ϵ.

Proof. Fix $\epsilon > 0$. Pick δ_A such that

$$\delta_A \leq \min \left\{ \frac{3\epsilon}{4C_0^2 \left(C_A + 1\right)}, \frac{\lambda}{C_P^2 + 4C_A + 4C_A C_P^2} \right\}$$

and choose $M\left(\delta_A\right)$ accordingly.

Let $v = u_N^\epsilon \in V_0$ in (5.15). Observe,

$$\begin{aligned} \mathcal{A}\left(u_N^\epsilon, u_N^\epsilon\right) &= \epsilon \left(\Delta u_N^\epsilon, \Delta u_N^\epsilon\right)_\Omega - \left(A : D^2 u_N^\epsilon, u_N^\epsilon\right)_\Omega \\ &= \epsilon \left\| \Delta u_N^\epsilon \right\|_{L^2(\Omega)}^2 - \left(A : D^2 u_N^\epsilon, u_N^\epsilon\right)_\Omega \end{aligned} \tag{5.17}$$

with

$$\begin{aligned} -\left(A : D^2 u_N^\epsilon, u_N^\epsilon\right)_\Omega &= -\sum_{j=1}^M \left((A - A_j + A_j) : D^2 u_N^\epsilon, u_N^\epsilon\right)_{\Omega_j} \\ &= -\sum_{j=1}^M \left((A - A_j) : D^2 u_N^\epsilon, u_N^\epsilon\right)_{\Omega_j} - \sum_{j=1}^M \left(A_j : D^2 u_N^\epsilon, u_N^\epsilon\right)_{\Omega_j} \\ &= -\sum_{j=1}^M \left((A - A_j) : D^2 u_N^\epsilon, u_N^\epsilon\right)_{\Omega_j} - \sum_{j=1}^M \left(\nabla \cdot A_j \nabla u_N^\epsilon, u_N^\epsilon\right)_{\Omega_j} \\ &= -\sum_{j=1}^M \left((A - A_j) : D^2 u_N^\epsilon, u_N^\epsilon\right)_{\Omega_j} - \sum_{j=1}^M \left(A_j \nabla u_N^\epsilon \cdot n_j, u_N^\epsilon\right)_{\partial\Omega_j} \\ &\quad + \sum_{j=1}^M \left(A_j \nabla u_N^\epsilon, \nabla u_N^\epsilon\right)_{\Omega_j}. \end{aligned} \tag{5.18}$$

Then,

$$\begin{aligned}
\sum_{j=1}^{M} \left((A-A_j) : D^2 u_N^\epsilon, u_N^\epsilon\right)_{\Omega_j} &\leq \sum_{j=1}^{M} \left| \left((A-A_j) : D^2 u_N^\epsilon, u_N^\epsilon\right)_{\Omega_j} \right| \qquad (5.19)\\
&\leq \sum_{j=1}^{M} \left\| (A-A_j) : D^2 u_N^\epsilon \right\|_{L^2(\Omega_j)} \left\| u_N^\epsilon \right\|_{L^2(\Omega_j)} \\
&\leq \sum_{j=1}^{M} \left\| |A-A_j| \left| D^2 u_N^\epsilon \right| \right\|_{L^2(\Omega_j)} \left\| u_N^\epsilon \right\|_{L^2(\Omega_j)} \\
&\leq \delta_A \sum_{j=1}^{M} \left\| D^2 u_N^\epsilon \right\|_{L^2(\Omega_j)} \left\| u_N^\epsilon \right\|_{L^2(\Omega_j)} \\
&\leq \delta_A \sum_{j=1}^{M} \left\| D^2 u_N^\epsilon \right\|^2_{L^2(\Omega_j)} + \frac{\delta_A}{4} \sum_{j=1}^{M} \left\| u_N^\epsilon \right\|^2_{L^2(\Omega_j)} \\
&= \delta_A \left\| D^2 u_N^\epsilon \right\|^2_{L^2(\Omega)} + \frac{\delta_A}{4} \left\| u_N^\epsilon \right\|^2_{L^2(\Omega)} \\
&\leq \delta_A C_0^2 \left\| \Delta u_N^\epsilon \right\|^2_{L^2(\Omega)} + \frac{\delta_A C_P^2}{4} \left\| \nabla u_N^\epsilon \right\|^2_{L^2(\Omega)},
\end{aligned}$$

$$\begin{aligned}
\sum_{j=1}^{M} \left(A_j \nabla u_N^\epsilon \cdot n_j, u_N^\epsilon\right)_{\partial\Omega_j} &\leq \delta_A C_A \left\| u_N^\epsilon \right\|^2_{H^2(\Omega)} \qquad (5.20)\\
&= \delta_A C_A \left\| D^2 u_N^\epsilon \right\|^2_{L^2(\Omega)} + \delta_A C_A \left\| \nabla u_N^\epsilon \right\|^2_{L^2(\Omega)} + \delta_A C_A \left\| u_N^\epsilon \right\|^2_{L^2(\Omega)} \\
&\leq \delta_A C_A C_0^2 \left\| \Delta u_N^\epsilon \right\|^2_{L^2(\Omega)} + \delta_A C_A \left(1 + C_P^2\right) \left\| \nabla u_N^\epsilon \right\|^2_{L^2(\Omega)}
\end{aligned}$$

and

$$\sum_{j=1}^{M} \left(A_j \nabla u_N^\epsilon, \nabla u_N^\epsilon\right)_{\Omega_j} \geq \sum_{j=1}^{M} \lambda \left\| \nabla u_N^\epsilon \right\|^2_{L^2(\Omega_j)} = \lambda \left\| \nabla u_N^\epsilon \right\|^2_{L^2(\Omega)}, \qquad (5.21)$$

where we have used Conjecture 5.1 for (5.20). Also,

$$
\begin{aligned}
(f, u_N^\epsilon)_\Omega &\leq |(f, u_N^\epsilon)_\Omega| \\
&\leq \|f\|_{L^2(\Omega)} \, \|u_N^\epsilon\|_{L^2(\Omega)} \\
&\leq \frac{C_P^2}{2\lambda} \|f\|_{L^2(\Omega)}^2 + \frac{\lambda}{2C_P^2} \|u_N^\epsilon\|_{L^2(\Omega)}^2 \\
&\leq \frac{C_P^2}{2\lambda} \|f\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|\nabla u_N^\epsilon\|_{L^2(\Omega)}^2 .
\end{aligned} \tag{5.22}
$$

Thus, combining relations (5.17) - (5.22), we have

$$
\begin{aligned}
&\epsilon \|\Delta u_N^\epsilon\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|\nabla u_N^\epsilon\|_{L^2(\Omega)}^2 \\
&\quad \leq \delta_A C_0^2 \left(C_A + 1\right) \|\Delta u_N^\epsilon\|_{L^2(\Omega)}^2 + \frac{\delta_A}{4} \left(C_P^2 + 4C_A + 4C_A C_P^2\right) \|\nabla u_N^\epsilon\|_{L^2(\Omega)}^2 \\
&\qquad + \frac{C_P^2}{2\lambda} \|f\|_{L^2(\Omega)}^2 .
\end{aligned} \tag{5.23}
$$

By the choice of δ_A, we have

$$
\frac{\epsilon}{4} \|\Delta u_N^\epsilon\|_{L^2(\Omega)}^2 + \frac{\lambda}{4} \|\nabla u_N^\epsilon\|_{L^2(\Omega)}^2 \leq \frac{C_P^2}{2\lambda} \|f\|_{L^2(\Omega)}^2 .
$$

Therefore,

$$
\epsilon \|\Delta u_N^\epsilon\|_{L^2(\Omega)}^2 + \lambda \|\nabla u_N^\epsilon\|_{L^2(\Omega)}^2 \leq \frac{2C_P^2}{\lambda} \|f\|_{L^2(\Omega)}^2 ,
$$

and we have

$$
\epsilon \left\| D^2 u_N^\epsilon \right\|_{L^2(\Omega)}^2 + \lambda C_0^2 \|\nabla u_N^\epsilon\|_{L^2(\Omega)}^2 \leq c,
$$

where $c = \frac{2C_P^2 C_0^2}{\lambda} \|f\|_{L^2(\Omega)}^2$. The proof is complete.

□

Using the a-priori estimate, we now show (5.15) has a unique solution. Fix a positive integer N. Then, we look for a function $u_N^\epsilon \in V_N$ such that

$$u_N^\epsilon = \sum_{j=1}^{N} \beta_N^j \phi_j \tag{5.24}$$

is a solution to (5.15) for some constants $\beta_N^1, \beta_N^2, \ldots, \beta_N^N \in \mathbb{R}$.

Proposition 5.1. *Problem* (5.15) *has a unique solution* $u_N^\epsilon \in V_N$.

Proof. Let $u_N^\epsilon = \sum_{j=1}^{N} \beta_N^j \phi_j$. Then, for $i = 1, 2, \ldots, N$, we have

$$\mathcal{A}\left(u_N^\epsilon, \phi_i\right) = \mathcal{A}\left(\sum_{j=1}^{N} \beta_N^j \phi_j, \phi_i\right) = \sum_{j=1}^{N} \beta_N^j \, \mathcal{A}\left(\phi_j, \phi_i\right). \tag{5.25}$$

Define $G \in \mathbb{R}^{N\times N}$, $\beta \in \mathbb{R}^N$, and $F \in \mathbb{R}^N$ by

$$[G]_{i,j} := \mathcal{A}\left(\phi_j, \phi_i\right), \qquad [\beta]_i := \beta_N^i, \qquad [F]_i := \left(f, \phi_i\right)_\Omega.$$

Then, $\mathcal{A}\left(u_N^\epsilon, v\right) = \left(f, v\right)_\Omega$ for all $v \in V_N$ is equivalent to

$$G\beta = F. \tag{5.26}$$

Thus, there exists constants $\beta_N^1, \beta_N^2, \ldots, \beta_N^N \in \mathbb{R}$ such that $u_N^\epsilon = \sum_{j=1}^{N} \beta_N^j \phi_j \in V_N$ is a solution to (5.15) if and only if G is invertible.

Let $f = 0$. Then $\left(f, \phi_i\right)_\Omega = 0$ for all $i = 1, 2, \ldots, N$, and we have $F = 0$. Suppose G is not invertible. Then, there exists $\beta \neq 0$ such that $G\beta = 0$, and we have, $\mathcal{A}\left(u_N^\epsilon, v\right) = 0$ for all $v \in V_N$. Thus, by Theorem 5.3,

$$\epsilon \left\|D^2 u_N^\epsilon\right\|_{L^2(\Omega)}^2 + \lambda\, C_0^2 \left\|\nabla u_N^\epsilon\right\|_{L^2(\Omega)}^2 \leq \frac{2C_P^2 C_0^2}{\lambda} \left\|f\right\|_{L^2(\Omega)}^2 = 0,$$

and we have u_N^ϵ must be constant. But, $u_N^\epsilon = 0$ on $\partial\Omega$ implies $u_N^\epsilon = 0$ on $\overline{\Omega}$, a contradiction. Therefore, G must be invertible. Hence, there exists a unique $\beta \in \mathbb{R}^N$

such that $u_N^\epsilon = \sum_{j=1}^N \beta_N^j \phi_j$ is a solution to (5.15), and it follows that (5.15) has a unique solution in V_N for all $N \geq 1$. The proof is complete. □

Finally, using the above two results, we are ready to prove the existence of a unique function $u^\epsilon \in V_0$ such that u^ϵ is a solution to (5.12) using a weak compactness argument.

Theorem 5.4. *For each $\epsilon > 0$, there exists a unique $u^\epsilon \in V_0$ such that*

$$\mathcal{A}(u^\epsilon, w) = (f, w)_\Omega$$

for all $w \in V_0$. Moreover,

$$\epsilon \left\| D^2 u^\epsilon \right\|_{L^2(\Omega)}^2 + \lambda c_1 \left\| \nabla u^\epsilon \right\|_{L^2(\Omega)}^2 \leq \frac{c_2}{\lambda} \left\| f \right\|_{L^2(\Omega)}^2, \tag{5.27}$$

where c_1, c_2 are positive constants which are independent of N, f, λ, and ϵ.

Proof. According to Theorem 5.3 and Theorem 5.2, we have the sequence $\{u_N^\epsilon\}_{N=1}^\infty$ is uniformly bounded in $V_0 = H^2(\Omega) \cap H_0^1(\Omega)$. Thus, there exists a subsequence $\left\{u_{N_m}^\epsilon\right\}_{m=1}^\infty \subset \{u_N^\epsilon\}_{N=1}^\infty$ and a function $u^\epsilon \in V_0$ such that $u_{N_m}^\epsilon \rightharpoonup u^\epsilon$ weakly in $H^2(\Omega)$, (cf. [22]).

Fix a positive integer M and choose a function $v \in V_0$ such that $v = \sum_{j=1}^M d_j \phi_j$, where $\{d_1, d_2, \ldots, d_M\}$ are given constants. Choose $m \geq M$. Then,

$$\mathcal{A}(u_m^\epsilon, v) = (f, v)_\Omega .$$

Set $m = N_m$. Then, taking the weak limit, we get

$$\mathcal{A}(u^\epsilon, v) = (f, v)_\Omega .$$

Since functions of the same form as v are dense in V_0, it follows that

$$\mathcal{A}(u^\epsilon, v) = (f, v)_\Omega .$$

for all $v \in V_0$. Thus, u^ϵ is a weak solution.

Suppose there are two functions u_1^ϵ and u_2^ϵ such that u_1^ϵ and u_2^ϵ are both solutions to (5.12). Let $u^\epsilon = u_1^\epsilon - u_2^\epsilon$. Then, by the linearity of $\mathcal{A}$,

$$\mathcal{A}(u^\epsilon, v) = 0 \qquad \forall v \in V_0.$$

Repeating the proof of Theorem 5.3, we can show that

$$\epsilon \left\| D^2 u^\epsilon \right\|_{L^2(\Omega)}^2 + \lambda C_0^2 \left\| \nabla u^\epsilon \right\|_{L^2(\Omega)}^2 \leq 0.$$

Therefore, $u^\epsilon \in V_0$ implies $u^\epsilon = 0$, and it follows that the solution to (5.12) is unique. The proof is complete.

□

5.2.3 Uniform H^2-Stability

So far we have shown there exists a unique function $u^\epsilon \in V_0$ such that u^ϵ is a solution to (5.12) with $\epsilon^{1/2} \left\| D^2 u^\epsilon \right\|_{L^2(\Omega)}$ uniformly bounded. Now we will show that $\left\| D^2 u^\epsilon \right\|_{L^2(\Omega)}$ is uniformly bounded independent of ϵ. To this end, we first need the following Lemma:

Lemma 5.1. *Let η_ρ denote a mollifier as in Section 5.2.1, and let $u_\rho^\epsilon = \eta_\rho * u^\epsilon$. Then there holds*

$$\begin{aligned} \frac{\epsilon}{4} \left\| \nabla \Delta u_\rho^\epsilon \right\|_{L^2(\Omega)}^2 + \frac{\lambda}{4} \left\| D^2 u_\rho^\epsilon \right\|_{L^2(\Omega)}^2 + \lambda \left\| \nabla u^\epsilon \right\|_{L^2(\Omega)}^2 - \frac{3\lambda}{4} \left\| \nabla u_\rho^\epsilon \right\|_{L^2(\Omega)}^2 \\ \leq \frac{1}{2\lambda} \left\| f_\rho \right\|_{L^2(\Omega)}^2 + \frac{2C_P^2}{\lambda} \left\| f \right\|_{L^2(\Omega)}^2 . \end{aligned}$$

Proof. Fix $\epsilon > 0$. Pick δ_A such that

$$\delta_A \leq \min\left\{ \frac{3\epsilon}{8C_A C_0^2}, \frac{\lambda}{8C_A + 10}, \frac{\epsilon}{C_0^2}, \frac{3\lambda}{8 + 8C_P^2} \right\} \tag{5.28}$$

and choose M accordingly.

Let $v \in V_0$ and $v_\rho = \eta_\rho * v$. Observe,

$$\begin{aligned} (f, v_\rho)_\Omega &= \int_{x\in\Omega} f(x) \int_{y\in\Omega} \eta_\rho(x-y) v(y)\, dydx \\ &= \int_{x\in\Omega}\int_{y\in\Omega} f(x)\eta_\rho(x-y)v(y)\, dydx \\ &= \int_{x\in\Omega}\int_{y\in\Omega} f(x)\eta_\rho(y-x)v(y)\, dydx \\ &= \int_{y\in\Omega} v(y) \int_{x\in\Omega} f(x)\eta_\rho(y-x)\, dxdy \\ &= (f_\rho, v)_\Omega\,, \end{aligned} \tag{5.29}$$

$$\epsilon\left(\Delta u^\epsilon, \Delta v_\rho\right)_\Omega = \epsilon\Big(\left(\Delta u^\epsilon\right)_\rho, \Delta v \Big)_\Omega = \epsilon\left(\Delta u^\epsilon_\rho, \Delta v\right)_\Omega, \tag{5.30}$$

and

$$\begin{aligned} \left(A : D^2 u^\epsilon, v_\rho\right)_\Omega &= \Big(\left(A : D^2 u^\epsilon\right)_\rho, v \Big)_\Omega \\ &= \sum_{j=1}^M \Big(\left[(A - A_j) : D^2 u^\epsilon\right]_\rho, v \Big)_{\Omega_j} + \sum_{j=1}^M \Big(\left(A_j : D^2 u^\epsilon\right)_\rho, v \Big)_{\Omega_j} \end{aligned} \tag{5.31}$$

with

$$\sum_{j=1}^{M}\Big(\left[(A-A_j):D^2u^\epsilon\right]_\rho,v\Big)_{\Omega_j}=\sum_{j=1}^{M}\Big(\eta_\rho*\left[(A-A_j):D^2u^\epsilon\right],v\Big)_{\Omega_j} \tag{5.32}$$
$$\leq\sum_{j=1}^{M}\|v\|_{L^2(\Omega_j)}\left\|\eta_\rho*\left[(A-A_j):D^2u^\epsilon\right]\right\|_{L^2(\Omega_j)}$$
$$\leq\sum_{j=1}^{M}\|v\|_{L^2(\Omega_j)}\|\eta_\rho\|_{L^\infty(\Omega_j)}\left\|\left[(A-A_j):D^2u^\epsilon\right]\right\|_{L^2(\Omega_j)}$$
$$\leq\sum_{j=1}^{M}\|v\|_{L^2(\Omega_j)}\left\|(A-A_j):D^2u^\epsilon\right\|_{L^2(\Omega_j)}$$
$$\leq\delta_A\sum_{j=1}^{M}\|v\|_{L^2(\Omega_j)}\left\|D^2u^\epsilon\right\|_{L^2(\Omega_j)}$$
$$\leq\frac{\delta_A}{2}\|v\|^2_{L^2(\Omega)}+\frac{\delta_A}{2}\left\|D^2u^\epsilon\right\|^2_{L^2(\Omega)},$$

$$\sum_{j=1}^{M}\Big(\left(A_j:D^2u^\epsilon\right)_\rho,v\Big)_{\Omega_j}=\sum_{j=1}^{M}\left(A_j:D^2u^\epsilon_\rho,v\right)_{\Omega_j} \tag{5.33}$$
$$=\sum_{j=1}^{M}\Big(\left(A_j-A\right):D^2u^\epsilon_\rho,v\Big)_{\Omega_j}+\sum_{j=1}^{M}\left(A:D^2u^\epsilon_\rho,v\right)_{\Omega_j}$$
$$=\sum_{j=1}^{M}\Big(\left(A_j-A\right):D^2u^\epsilon_\rho,v\Big)_{\Omega_j}+\left(A:D^2u^\epsilon_\rho,v\right)_{\Omega},$$

and

$$\sum_{j=1}^{M}\Big(\left(A_j-A\right):D^2u^\epsilon_\rho,v\Big)_{\Omega_j}\leq\delta_A\sum_{j=1}^{M}\left\|D^2u^\epsilon_\rho\right\|_{L^2(\Omega_j)}\|v\|_{L^2(\Omega_j)} \tag{5.34}$$
$$\leq\frac{\delta_A}{2}\left\|D^2u^\epsilon_\rho\right\|^2_{L^2(\Omega)}+\frac{\delta_A}{2}\|v\|^2_{L^2(\Omega)}.$$

Thus, combining relations (5.29) - (5.34), we have

$$\mathcal{A}\left(u^\epsilon_\rho,v\right)\leq\left|\left(f_\rho,v\right)_\Omega\right|+\delta_A\|v\|^2_{L^2(\Omega)}+\frac{\delta_A}{2}\left\|D^2u^\epsilon\right\|^2_{L^2(\Omega)}+\frac{\delta_A}{2}\left\|D^2u^\epsilon_\rho\right\|^2_{L^2(\Omega)} \tag{5.35}$$

for all $v \in V_0$.

Let $v = -\Delta u_\rho^\epsilon \in V_0$ in (5.35). We have

$$\begin{aligned} -\mathcal{A}\left(u_\rho^\epsilon, \Delta u_\rho^\epsilon\right) &= -\epsilon\left(\Delta u_\rho^\epsilon, \Delta^2 u_\rho^\epsilon\right)_\Omega + \left(A : D^2 u_\rho^\epsilon, \Delta u_\rho^\epsilon\right)_\Omega \\ &= \epsilon\left(\nabla\Delta u_\rho^\epsilon, \nabla\Delta u_\rho^\epsilon\right)_\Omega + \left(A : D^2 u_\rho^\epsilon, \Delta u_\rho^\epsilon\right)_\Omega \\ &= \epsilon\left\|\nabla\Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \left(A : D^2 u_\rho^\epsilon, \Delta u_\rho^\epsilon\right)_\Omega \end{aligned} \tag{5.36}$$

and

$$\begin{aligned} \left(A : D^2 u_\rho^\epsilon, \Delta u_\rho^\epsilon\right)_\Omega &= \sum_{j=1}^M \Big((A_j - A_j + A) : D^2 u_\rho^\epsilon, \Delta u_\rho^\epsilon \Big)_{\Omega_j} \\ &= \sum_{j=1}^M \left(A_j : D^2 u_\rho^\epsilon, \Delta u_\rho^\epsilon\right)_{\Omega_j} + \sum_{j=1}^M \Big((A - A_j) : D^2 u_\rho^\epsilon, \Delta u_\rho^\epsilon \Big)_{\Omega_j}, \end{aligned} \tag{5.37}$$

with

$$\begin{aligned} \sum_{j=1}^M \left(A_j : D^2 u_\rho^\epsilon, \Delta u_\rho^\epsilon\right)_{\Omega_j} &= \sum_{j=1}^M \left(A_j \nabla u_\rho^\epsilon \cdot n_j, \Delta u_\rho^\epsilon\right)_{\partial\Omega_j} \\ &\quad - \sum_{j=1}^M \left(A_j \nabla u_\rho^\epsilon, D^2 u_\rho^\epsilon n_j\right)_{\partial\Omega_j} + \sum_{j=1}^M \left(A_j D^2 u_\rho^\epsilon, D^2 u_\rho^\epsilon\right)_{\Omega_j}. \end{aligned} \tag{5.38}$$

Then,

$$\begin{aligned} \sum_{j=1}^M \left(A_j D^2 u_\rho^\epsilon, D^2 u_\rho^\epsilon\right)_{\Omega_j} &= \sum_{j=1}^M \sum_{k=1}^n \Big(A_j \left[D^2 u_\rho^\epsilon\right]_k, \left[D^2 u_\rho^\epsilon\right]_k \Big)_{\Omega_j} \\ &\geq \lambda \sum_{j=1}^M \sum_{k=1}^n \Big(\left[D^2 u_\rho^\epsilon\right]_k, \left[D^2 u_\rho^\epsilon\right]_k \Big)_{\Omega_j} \\ &= \lambda \sum_{j=1}^M \left(D^2 u_\rho^\epsilon, D^2 u_\rho^\epsilon\right)_{\Omega_j} \\ &= \lambda \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2, \end{aligned} \tag{5.39}$$

$$\sum_{j=1}^{M} \left(A_j \nabla u_\rho^\epsilon, D^2 u_\rho^\epsilon n_j\right)_{\partial\Omega_j} \le \delta_A C_A \left\|u_\rho^\epsilon\right\|_{H^3(\Omega)}^2 \tag{5.40}$$

$$\begin{aligned} &= \delta_A C_A \left(\left\|D^3 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2\right) \\ &\quad + \delta_A C_A \left(\left\|\nabla u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \left\|u_\rho^\epsilon\right\|_{L^2(\Omega)}^2\right) \\ &\le \delta_A C_A \left(C_0^2 \left\|\nabla \Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2\right) \\ &\quad + \delta_A C_A \left(1 + C_P^2\right) \left\|\nabla u_\rho^\epsilon\right\|_{L^2(\Omega)}^2, \end{aligned}$$

$$-\sum_{j=1}^{M} \left(A_j \nabla u_\rho^\epsilon \cdot n_j, \Delta u_\rho^\epsilon\right)_{\partial\Omega_j} \le \delta_A C_A \left\|u_\rho^\epsilon\right\|_{H^3(\Omega)}^2 \tag{5.41}$$

$$\begin{aligned} &\le \delta_A C_A \left(C_0^2 \left\|\nabla \Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2\right) \\ &\quad + \delta_A C_A \left(1 + C_P^2\right) \left\|\nabla u_\rho^\epsilon\right\|_{L^2(\Omega)}^2, \end{aligned}$$

and

$$-\sum_{j=1}^{M} \Big((A - A_j) : D^2 u_\rho^\epsilon, \Delta u_\rho^\epsilon \Big)_{\Omega_j} \le \delta_A \sum_{j=1}^{M} \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega_j)} \left\|\Delta u_\rho^\epsilon\right\|_{L^2(\Omega_j)} \tag{5.42}$$

$$\begin{aligned} &\le \frac{\delta_A}{2} \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\delta_A}{2} \left\|\Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 \\ &\le \delta_A \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2, \end{aligned}$$

where we have used Conjecture 5.1 to show (5.40) and (5.41). Also,

$$\left|\left(f_\rho, \Delta u_\rho^\epsilon\right)_\Omega\right| \le \|f_\rho\|_{L^2(\Omega)} \left\|\Delta u_\rho^\epsilon\right\|_{L^2(\Omega)} \tag{5.43}$$

$$\begin{aligned} &\le \frac{1}{2\lambda} \|f_\rho\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \left\|\Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 \\ &\le \frac{1}{2\lambda} \|f_\rho\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2. \end{aligned}$$

Thus, plugging relations (5.36) - (5.43) into (5.35), we have

$$\begin{aligned}
&\epsilon \left\|\nabla \Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 \\
&\quad \leq 2\delta_A C_A C_0^2 \left\|\nabla \Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\delta_A}{2}\left(4C_A + 5\right) \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 \\
&\qquad + \frac{\delta_A}{2} \left\|D^2 u^\epsilon\right\|_{L^2(\Omega)}^2 + 2\delta_A \left(1 + C_P^2\right) \left\|\nabla u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{1}{2\lambda} \left\|f_\rho\right\|_{L^2(\Omega)}^2 .
\end{aligned} \tag{5.44}$$

Combining (5.44) and (5.27), and using the bounds for δ_A, we have

$$\begin{aligned}
&\epsilon \left\|\nabla \Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\epsilon}{C_0^2} \left\|D^2 u^\epsilon\right\|_{L^2(\Omega)}^2 + \lambda \left\|\nabla u^\epsilon\right\|_{L^2(\Omega)}^2 \\
&\quad \leq 2\delta_A C_A C_0^2 \left\|\nabla \Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\delta_A}{2}\left(4C_A + 5\right) \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 \\
&\qquad + \frac{\delta_A}{2} \left\|D^2 u^\epsilon\right\|_{L^2(\Omega)}^2 + 2\delta_A \left(1 + C_P^2\right) \left\|\nabla u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{1}{2\lambda} \left\|f_\rho\right\|_{L^2(\Omega)}^2 + \frac{2C_P^2}{\lambda} \left\|f\right\|_{L^2(\Omega)}^2 \\
&\quad \leq \frac{3\epsilon}{4} \left\|\nabla \Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\lambda}{4} \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\epsilon}{2C_0^2} \left\|D^2 u^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{3\lambda}{4} \left\|\nabla u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 \\
&\qquad + \frac{1}{2\lambda} \left\|f_\rho\right\|_{L^2(\Omega)}^2 + \frac{2C_P^2}{\lambda} \left\|f\right\|_{L^2(\Omega)}^2 .
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{\epsilon}{4} \left\|\nabla \Delta u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\lambda}{4} \left\|D^2 u_\rho^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\epsilon}{2C_0^2} \left\|D^2 u^\epsilon\right\|_{L^2(\Omega)}^2 + \lambda \left\|\nabla u^\epsilon\right\|_{L^2(\Omega)}^2 \\
&\quad \leq \frac{1}{2\lambda} \left\|f_\rho\right\|_{L^2(\Omega)}^2 + \frac{2C_P^2}{\lambda} \left\|f\right\|_{L^2(\Omega)}^2 + \frac{3\lambda}{4} \left\|\nabla u_\rho^\epsilon\right\|_{L^2(\Omega)}^2
\end{aligned}$$

and the result follows since $\frac{\epsilon}{2C_0^2} \left\|D^2 u^\epsilon\right\|_{L^2(\Omega)}^2 > 0$. The proof is complete. □

Using Lemma 5.1, we can now show the following improved stability estimate.

Theorem 5.5. *Let $u^\epsilon \in V_0$ be the unique solution of (5.12), then*

$$\left\|D^2 u^\epsilon\right\|_{L^2(\Omega)}^2 + \left\|\nabla u^\epsilon\right\|_{L^2(\Omega)}^2 \leq \frac{c_1}{\lambda^2} \left\|f\right\|_{L^2(\Omega)}^2 , \tag{5.45}$$

and, if $u^\epsilon \in H^3(\Omega)$,

$$\epsilon \left\| D^3 u^\epsilon \right\|^2_{L^2(\Omega)} + \lambda\, c_2 \left\| D^2 u^\epsilon \right\|^2_{L^2(\Omega)} + \lambda\, c_2 \left\| \nabla u^\epsilon \right\|^2_{L^2(\Omega)} \leq \frac{c_3}{\lambda} \left\| f \right\|^2_{L^2(\Omega)}, \tag{5.46}$$

where c_1, c_2, *and* c_3 *are positive constants which are independent of* f, λ, *and* ϵ.

Proof. By Lemma 5.1, we have

$$\frac{\lambda}{4} \left\| D^2 u^\epsilon_\rho \right\|^2_{L^2(\Omega)} + \lambda \left\| \nabla u^\epsilon \right\|^2_{L^2(\Omega)} - \frac{3\lambda}{4} \left\| \nabla u^\epsilon_\rho \right\|^2_{L^2(\Omega)} \leq \frac{1}{2\lambda} \left\| f_\rho \right\|^2_{L^2(\Omega)} + \frac{2C_P^2}{\lambda} \left\| f \right\|^2_{L^2(\Omega)}.$$

Letting $\rho \to 0$, we have

$$\frac{\lambda}{4} \left\| D^2 u^\epsilon \right\|^2_{L^2(\Omega)} + \frac{\lambda}{4} \left\| \nabla u^\epsilon \right\|^2_{L^2(\Omega)} \leq \frac{4C_P^2 + 1}{2\lambda} \left\| f \right\|^2_{L^2(\Omega)}$$

since $u^\epsilon \in H^2(\Omega)$ and $f \in L^2(\Omega)$. Thus,

$$\left\| D^2 u^\epsilon \right\|^2_{L^2(\Omega)} + \left\| \nabla u^\epsilon \right\|^2_{L^2(\Omega)} \leq c,$$

where $c = \frac{8C_P^2+2}{\lambda^2} \left\| f \right\|^2_{L^2(\Omega)}$.

Similarly, if $u^\epsilon \in H^3(\Omega)$, letting $\rho \to 0$ in Lemma 5.1 gives

$$\epsilon \left\| \nabla \Delta u^\epsilon \right\|^2_{L^2(\Omega)} + \lambda \left\| D^2 u^\epsilon \right\|^2_{L^2(\Omega)} + \lambda \left\| \nabla u^\epsilon \right\|^2_{L^2(\Omega)} \leq \frac{8C_P^2 + 2}{\lambda} \left\| f \right\|^2_{L^2(\Omega)}.$$

Thus, by Theorem 5.2,

$$\epsilon \left\| D^3 u^\epsilon \right\|^2_{L^2(\Omega)} + \lambda C_0^2 \left\| D^2 u^\epsilon \right\|^2_{L^2(\Omega)} + \lambda C_0^2 \left\| \nabla u^\epsilon \right\|^2_{L^2(\Omega)} \leq c,$$

where $c = \frac{\left(8C_P^2+2\right)C_0^2}{\lambda} \left\| f \right\|^2_{L^2(\Omega)}$. The proof is complete.

□

5.2.4 Convergence

Using the results of the previous two sections, we can now show that u^ϵ converges to the strong solution u of problem (5.8) as $\epsilon \to 0$. Furthermore, we can derive convergence rates in various H^p norms in terms of $O(\epsilon^r)$ for $p = 0, 1, 2$ and $0 < r \leq 1$. Recall, the convergence is necessary in order to apply the vanishing moment methodology to problem (5.8).

In order to derive the L^2 and H^1 rates of convergence, we will use a duality argument (see [6]). To this end, we first derive an approximate adjoint operator for L. Pick a positive integer M and cover Ω with a uniform grid such that M cells cover Ω. Let $v, w \in V_0$. Observe,

$$
\begin{aligned}
-(Lv, w)_\Omega &= \left(A : D^2 v, w\right)_\Omega \\
&= \Big((A^s + A^a) : D^2 v, w \Big)_\Omega \\
&= \left(A^s : D^2 v, w\right)_\Omega + \left(A^a : D^2 v, w\right)_\Omega \\
&= \sum_{j=1}^{M} \Big(\left(A^s - A^s_j\right) : D^2 v, w \Big)_{\Omega_j} + \sum_{j=1}^{M} \left(A^s_j : D^2 v, w\right)_{\Omega_j} \\
&\quad + \sum_{j=1}^{M} \Big(\left(A^a - A^a_j\right) : D^2 v, w \Big)_{\Omega_j} + \sum_{j=1}^{M} \left(A^a_j : D^2 v, w\right)_{\Omega_j}
\end{aligned}
\tag{5.47}
$$

with

$$\begin{aligned}
\sum_{j=1}^{M} \left(A_j^s : D^2 v, w\right)_{\Omega_j} &= \sum_{j=1}^{M} \left(\nabla \cdot A_j^s \nabla v, w\right)_{\Omega_j} \\
&= \sum_{j=1}^{M} \left(A_j^s \nabla v \cdot n_j, w\right)_{\partial\Omega_j} - \sum_{j=1}^{M} \left(A_j^s \nabla v, \nabla w\right)_{\Omega_j} \\
&= \sum_{j=1}^{M} \left(A_j^s \nabla v \cdot n_j, w\right)_{\partial\Omega_j} - \sum_{j=1}^{M} \left(\nabla v, A_j^s \nabla w\right)_{\Omega_j} \\
&= \sum_{j=1}^{M} \left(A_j^s \nabla v \cdot n_j, w\right)_{\partial\Omega_j} - \sum_{j=1}^{M} \left(v, A_j^s \nabla w \cdot n_j\right)_{\partial\Omega_j} \\
&\quad + \sum_{j=1}^{M} \left(v, \nabla \cdot A_j^s \nabla w\right)_{\Omega_j} \\
&= \sum_{j=1}^{M} \left(A_j^s \nabla v \cdot n_j, w\right)_{\partial\Omega_j} - \sum_{j=1}^{M} \left(v, A_j^s \nabla w \cdot n_j\right)_{\partial\Omega_j} \\
&\quad + \sum_{j=1}^{M} \left(v, A_j^s : D^2 w\right)_{\Omega_j} \\
&= \sum_{j=1}^{M} \left(A_j^s \nabla v \cdot n_j, w\right)_{\partial\Omega_j} - \sum_{j=1}^{M} \left(v, A_j^s \nabla w \cdot n_j\right)_{\partial\Omega_j} \\
&\quad + \sum_{j=1}^{M} \Big(v, \left(A_j^s - A^s\right) : D^2 w\Big)_{\Omega_j} + \left(v, A^s : D^2 w\right)_{\Omega}
\end{aligned} \tag{5.48}$$

and, similarly,

$$\begin{aligned}
\sum_{j=1}^{M} \left(A_j^a : D^2 v, w\right)_{\Omega_j} &= \sum_{j=1}^{M} \left(A_j^a \nabla v \cdot n_j, w\right)_{\partial\Omega_j} + \sum_{j=1}^{M} \left(v, A_j^a \nabla w \cdot n_j\right)_{\partial\Omega_j} \\
&\quad - \sum_{j=1}^{M} \Big(v, \left(A_j^a - A^a\right) : D^2 w\Big)_{\Omega_j} - \left(v, A^a : D^2 w\right)_{\Omega}.
\end{aligned} \tag{5.49}$$

Therefore, combining relations (5.47) - (5.49), we have

$$
\begin{aligned}
-\left(Lv, w\right)_{\Omega} = & \Big(v, \left(A^s - A^a\right) : D^2 w\Big)_{\Omega} + \sum_{j=1}^{M} \Big(\left(A^s - A_j^s\right) : D^2 v, w\Big)_{\Omega_j} \\
& + \sum_{j=1}^{M} \Big(\left(A^a - A_j^a\right) : D^2 v, w\Big)_{\Omega_j} + \sum_{j=1}^{M} \left(A_j \nabla v \cdot n_j, w\right)_{\partial\Omega_j} \\
& + \sum_{j=1}^{M} \Big(v, \left(A_j^a - A_j^s\right) \nabla w \cdot n_j\Big)_{\partial\Omega_j} + \sum_{j=1}^{M} \Big(v, \left(A_j^s - A^s\right) : D^2 w\Big)_{\Omega_j} \\
& - \sum_{j=1}^{M} \Big(v, \left(A_j^a - A^a\right) : D^2 w\Big)_{\Omega_j} \\
= & -\Big(v, L^* w\Big)_{\Omega} + \sum_{j=1}^{M} \Big(\left(A^s - A_j^s\right) : D^2 v, w\Big)_{\Omega_j} \\
& + \sum_{j=1}^{M} \Big(\left(A^a - A_j^a\right) : D^2 v, w\Big)_{\Omega_j} + \sum_{j=1}^{M} \left(A_j \nabla v \cdot n_j, w\right)_{\partial\Omega_j} \\
& + \sum_{j=1}^{M} \Big(v, \left(A_j^a - A_j^s\right) \nabla w \cdot n_j\Big)_{\partial\Omega_j} + \sum_{j=1}^{M} \Big(v, \left(A_j^s - A^s\right) : D^2 w\Big)_{\Omega_j} \\
& - \sum_{j=1}^{M} \Big(v, \left(A_j^a - A^a\right) : D^2 w\Big)_{\Omega_j}
\end{aligned}
\tag{5.50}
$$

for all $v, w \in V_0$, where we have defined our approximate adjoint operator L^* by

$$
L^* v := -\left(A^s - A^a\right) : D^2 v \tag{5.51}
$$

for all $v \in V_0$.

To apply the duality argument, we need an error equation and a dual problem involving the error. Notice, if $u \in V_g$ is the strong solution of (5.8), then

$$
-\left(A : D^2 u, v\right)_{\Omega} = \left(f, v\right)_{\Omega} \tag{5.52}
$$

for all $v \in V_0$. Define $e^\epsilon \in V_0$ by

$$e^\epsilon := u^\epsilon - u.$$

Then, subtracting (5.52) from (5.12), we have the error equation

$$\left(A : D^2 e^\epsilon, v\right)_\Omega = \epsilon \left(\Delta u^\epsilon, \Delta v\right)_\Omega \tag{5.53}$$

for all $v \in V_0$. We can also define a dual problem for the error by

$$L^* w = e^\epsilon \qquad \text{in } \Omega, \tag{5.54a}$$

$$w = 0 \qquad \text{on } \partial\Omega. \tag{5.54b}$$

Since $A^s - A^a$ satisfies the same ellipticity condition as A, by [32], we have there exists a unique solution $w \in V_0$ to (5.54) with

$$\|w\|_{H^2(\Omega)} \leq C_D \|L^* w\|_{L^2(\Omega)} = C_D \|e^\epsilon\|_{L^2(\Omega)} \tag{5.55}$$

for some constant $C_D > 0$. Using the above observations, we are now ready to use a duality argument to derive L^2 and H^1 convergence rates.

Theorem 5.6. *Let $u^\epsilon \in V_g$ be the unique solution of (5.12). If $u \in V_g$ is the strong solution of (5.8), then*

$$\|u - u^\epsilon\|_{L^2(\Omega)} \leq \epsilon \left(\frac{c_1}{\lambda^2} \|f\|^2_{L^2(\Omega)} + c_2 \|u\|^2_{H^2(\Omega)}\right)^{1/2}, \tag{5.56}$$

$$\|\nabla(u - u^\epsilon)\|_{L^2(\Omega)} \leq \sqrt{\epsilon} \left(\frac{c_3}{\lambda^3} \|f\|^2_{L^2(\Omega)} + \frac{c_4}{\lambda} \left\|D^2 u\right\|^2_{L^2(\Omega)}\right)^{1/2}, \tag{5.57}$$

where c_1, c_2, c_3, and c_4 are positive constants which are independent of f, λ, u, and ϵ.

Proof. Fix $\epsilon > 0$. Since A, A^s, and A^a are uniformly continuous on $\overline{\Omega}$, we now define $M = M(\delta_A)$ such that, for the given partition, if $x, y \in \Omega_j$ for some $j = 1, 2, \ldots, M$, then $|A(x) - A(y)| < \delta_A$, $|A^s(x) - A^s(y)| < \delta_A$, and $|A^a(x) - A^a(y)| < \delta_A$, where the constant $\delta_A > 0$ is chosen such that

$$\delta_A \leq \min\left\{\frac{\lambda}{4C_A(1+C_P^2)}, \frac{\lambda}{2C_P^2}, \frac{\epsilon}{2(1+C_A)}, \frac{\epsilon^2 C_D^2}{2+3C_B}, \frac{1}{4(2+4C_D^2+3C_B C_D^2)}\right\}.$$

Let $v = -e^\epsilon \in V_0$ in (5.53). Observe,

$$-\left(A : D^2 e^\epsilon, e^\epsilon\right)_\Omega = -\sum_{j=1}^M \left((A - A_j) : D^2 e^\epsilon, e^\epsilon\right)_{\Omega_j} - \sum_{j=1}^M \left(A_j : D^2 e^\epsilon, e^\epsilon\right)_{\Omega_j}. \quad (5.58)$$

Then,

$$\begin{aligned}\sum_{j=1}^M \left((A - A_j) : D^2 e^\epsilon, e^\epsilon\right)_{\Omega_j} &\leq \delta_A \sum_{j=1}^M \left\|D^2 e^\epsilon\right\|_{L^2(\Omega_j)} \left\|e^\epsilon\right\|_{L^2(\Omega_j)} \\ &\leq \frac{\delta_A}{2}\left\|D^2 e^\epsilon\right\|^2_{L^2(\Omega)} + \frac{\delta_A}{2}\left\|e^\epsilon\right\|^2_{L^2(\Omega)} \\ &\leq \delta_A \left\|D^2 u^\epsilon\right\|^2_{L^2(\Omega)} + \delta_A \left\|D^2 u\right\|^2_{L^2(\Omega)} \\ &\quad + \frac{\delta_A C_P^2}{2}\left\|\nabla e^\epsilon\right\|^2_{L^2(\Omega)}\end{aligned} \quad (5.59)$$

and

$$-\sum_{j=1}^M \left(A_j : D^2 e^\epsilon, e^\epsilon\right)_{\Omega_j} = -\sum_{j=1}^M \left(A_j \nabla e^\epsilon \cdot n_j, e^\epsilon\right)_{\partial\Omega_j} + \sum_{j=1}^M \left(A_j \nabla e^\epsilon, \nabla e^\epsilon\right)_{\Omega_j}, \quad (5.60)$$

with

$$\sum_{j=1}^{M} \left(A_j \nabla e^\epsilon \cdot n_j, e^\epsilon\right)_{\partial\Omega_j} \leq \delta_A C_A \left\|e^\epsilon\right\|^2_{H^2(\Omega)} \tag{5.61}$$

$$\leq \delta_A C_A \left\|D^2 e^\epsilon\right\|^2_{L^2(\Omega)} + \delta_A C_A (1 + C_P^2) \left\|\nabla e^\epsilon\right\|^2_{L^2(\Omega)}$$

$$\leq \delta_A C_A \left(\left\|D^2 u^\epsilon\right\|^2_{L^2(\Omega)} + \left\|D^2 u\right\|^2_{L^2(\Omega)}\right)$$

$$+ \delta_A C_A (1 + C_P^2) \left\|\nabla e^\epsilon\right\|^2_{L^2(\Omega)},$$

$$\sum_{j=1}^{M} \left(A_j \nabla e^\epsilon, \nabla e^\epsilon\right)_{\Omega_j} \geq \lambda \sum_{j=1}^{M} \left\|\nabla e^\epsilon\right\|^2_{L^2(\Omega_j)} = \lambda \left\|\nabla e^\epsilon\right\|^2_{L^2(\Omega)}, \tag{5.62}$$

and

$$-\epsilon \left(\Delta u^\epsilon, \Delta e^\epsilon\right)_\Omega \leq \frac{\epsilon}{2} \left\|\Delta u^\epsilon\right\|^2_{L^2(\Omega)} + \frac{\epsilon}{2} \left\|\Delta e^\epsilon\right\|^2_{L^2(\Omega)} \tag{5.63}$$

$$\leq \frac{\epsilon}{2} \left\|D^2 u^\epsilon\right\|^2_{L^2(\Omega)} + \frac{\epsilon}{2} \left\|D^2 u\right\|^2_{L^2(\Omega)},$$

where we have used Conjecture 5.1 to show (5.61). Thus, combining relations (5.58) - (5.63) and using the inequality for δ_A, we have

$$\left\|\nabla e^\epsilon\right\|^2_{L^2(\Omega)} \leq \frac{2\epsilon}{\lambda} \left(\left\|D^2 u^\epsilon\right\|^2_{L^2(\Omega)} + \left\|D^2 u\right\|^2_{L^2(\Omega)}\right). \tag{5.64}$$

Hence, by Theorem 5.5,

$$\left\|\nabla e^\epsilon\right\|^2_{L^2(\Omega)} \leq \frac{\epsilon}{\lambda} \left[\frac{16 C_P^2 + 4}{\lambda^2} \left\|f\right\|^2_{L^2(\Omega)} + 2 \left\|D^2 u\right\|^2_{L^2(\Omega)}\right],$$

and we have the desired bound for $\left\|\nabla e^\epsilon\right\|_{L^2(\Omega)}$.

Now, we show $\|e^\epsilon\|_{L^2(\Omega)} = O(\epsilon)$ using a duality argument and Conjecture 5.2. Using (5.50) and (5.54), we have

$$\begin{aligned}
\|e^\epsilon\|^2_{L^2(\Omega)} &= (e^\epsilon, e^\epsilon)_\Omega \\
&= (e^\epsilon, L^* w)_\Omega \\
&= (Le^\epsilon, w)_\Omega + \sum_{j=1}^{M} \Big(\left(A^s - A^s_j\right) : D^2 e^\epsilon, w \Big)_{\Omega_j} \\
&\quad + \sum_{j=1}^{M} \Big(\left(A^a - A^a_j\right) : D^2 e^\epsilon, w \Big)_{\Omega_j} + \sum_{j=1}^{M} (A_j \nabla e^\epsilon \cdot n_j, w)_{\partial\Omega_j} \\
&\quad + \sum_{j=1}^{M} \Big(e^\epsilon, \left(A^a_j - A^s_j\right) \nabla w \cdot n_j \Big)_{\partial\Omega_j} + \sum_{j=1}^{M} \Big(e^\epsilon, \left(A^s_j - A^s\right) : D^2 w \Big)_{\Omega_j} \\
&\quad - \sum_{j=1}^{M} \Big(e^\epsilon, \left(A^a_j - A^a\right) : D^2 w \Big)_{\Omega_j}.
\end{aligned} \tag{5.65}$$

Then, by (5.53) and (5.55), we have

$$\begin{aligned}
(Le^\epsilon, w)_\Omega &= -\epsilon \left(\Delta u^\epsilon, \Delta w\right)_\Omega \\
&\leq \epsilon \left\|\Delta u^\epsilon\right\|_{L^2(\Omega)} \left\|\Delta w\right\|_{L^2(\Omega)} \\
&\leq \epsilon \left\|D^2 u^\epsilon\right\|_{L^2(\Omega)} \left\|w\right\|_{H^2(\Omega)} \\
&\leq \epsilon C_D \left\|D^2 u^\epsilon\right\|_{L^2(\Omega)} \left\|e^\epsilon\right\|_{L^2(\Omega)} \\
&\leq \epsilon^2 C_D^2 \left\|D^2 u^\epsilon\right\|^2_{L^2(\Omega)} + \frac{1}{2} \left\|e^\epsilon\right\|^2_{L^2(\Omega)}.
\end{aligned} \tag{5.66}$$

Also,

$$\sum_{j=1}^{M}\Big(\left(A^s-A^s_j\right):D^2e^\epsilon,w\Big)_{\Omega_j}\le\delta_A\left\|D^2e^\epsilon\right\|^2_{L^2(\Omega)}+\delta_A\left\|w\right\|^2_{L^2(\Omega)} \tag{5.67}$$

$$\le\delta_A\left(\left\|D^2u^\epsilon\right\|^2_{L^2(\Omega)}+\left\|D^2u\right\|^2_{L^2(\Omega)}\right)$$

$$+\delta_A\left\|w\right\|^2_{H^2(\Omega)}$$

$$\le\delta_A\left(\left\|D^2u^\epsilon\right\|^2_{L^2(\Omega)}+\left\|D^2u\right\|^2_{L^2(\Omega)}\right)$$

$$+\delta_AC_D^2\left\|e^\epsilon\right\|^2_{L^2(\Omega)},$$

$$\sum_{j=1}^{M}\Big(\left(A^a-A^a_j\right):D^2e^\epsilon,w\Big)_{\Omega_j}\le\delta_A\left(\left\|D^2u^\epsilon\right\|^2_{L^2(\Omega)}+\left\|D^2u\right\|^2_{L^2(\Omega)}\right) \tag{5.68}$$

$$+\delta_AC_D^2\left\|e^\epsilon\right\|^2_{L^2(\Omega)},$$

and

$$\sum_{j=1}^{M}\left(A_j\nabla e^\epsilon\cdot n_j,w\right)_{\partial\Omega_j}\le\delta_AC_B\left\|e^\epsilon\right\|^2_{H^2(\Omega)}+\delta_AC_B\left\|w\right\|^2_{H^2(\Omega)} \tag{5.69}$$

$$\le\delta_AC_B\left\|u^\epsilon\right\|^2_{H^2(\Omega)}+\delta_AC_B\left\|u\right\|^2_{H^2(\Omega)}$$

$$+\delta_AC_BC_D^2\left\|e^\epsilon\right\|^2_{L^2(\Omega)},$$

where we have used Conjecture 5.2 to show (5.69). Similarly,

$$\begin{aligned}\sum_{j=1}^{M}\left(e^{\epsilon},\left(A_j^a-A_j^s\right)\nabla w\cdot n_j\right)_{\partial\Omega_j} &= \sum_{j=1}^{M}\left(e^{\epsilon},A_j^a\nabla w\cdot n_j\right)_{\partial\Omega_j} \\ &\quad -\sum_{j=1}^{M}\left(e^{\epsilon},A_j^s\nabla w\cdot n_j\right)_{\partial\Omega_j} \\ &\leq 2\delta_A C_B\left\|e^{\epsilon}\right\|_{H^2(\Omega)}^2+2\delta_A C_B\left\|w\right\|_{H^2(\Omega)}^2 \\ &\leq 2\delta_A C_B\left(\left\|u^{\epsilon}\right\|_{H^2(\Omega)}^2+\left\|u\right\|_{H^2(\Omega)}^2\right) \\ &\quad +2\delta_A C_B C_D^2\left\|e^{\epsilon}\right\|_{L^2(\Omega)}^2,\end{aligned} \tag{5.70}$$

$$\begin{aligned}\sum_{j=1}^{M}\left(e^{\epsilon},\left(A_j^s-A^s\right):D^2w\right)_{\Omega_j} &\leq \delta_A\left\|e^{\epsilon}\right\|_{L^2(\Omega)}^2+\delta_A\left\|D^2w\right\|_{L^2(\Omega)}^2 \\ &\leq \delta_A\left\|e^{\epsilon}\right\|_{L^2(\Omega)}^2+\delta_A\left\|w\right\|_{H^2(\Omega)}^2 \\ &\leq \delta_A\left(1+C_D^2\right)\left\|e^{\epsilon}\right\|_{L^2(\Omega)}^2,\end{aligned} \tag{5.71}$$

and

$$-\sum_{j=1}^{M}\left(e^{\epsilon},\left(A_j^a-A^a\right):D^2w\right)_{\Omega_j}\leq\delta_A\left(1+C_D^2\right)\left\|e^{\epsilon}\right\|_{L^2(\Omega)}^2, \tag{5.72}$$

where we have used Conjecture 5.2 to shown (5.70). Thus, combining (5.65) - (5.72), we have

$$\begin{aligned}\frac{1}{2}\left\|e^{\epsilon}\right\|_{L^2(\Omega)}^2 &\leq \left(\epsilon^2C_D^2+2\delta_A+3C_B\delta_A\right)\left\|u^{\epsilon}\right\|_{H^2(\Omega)}^2+\delta_A\left(2+3C_B\right)\left\|u\right\|_{H^2(\Omega)}^2 \\ &\quad +\delta_A\left(2+4C_D^2+3C_BC_D^2\right)\left\|e^{\epsilon}\right\|_{L^2(\Omega)}^2,\end{aligned} \tag{5.73}$$

and by the choice of δ_A, it follows that

$$\left\|e^{\epsilon}\right\|_{L^2(\Omega)}^2\leq\epsilon^2\,8C_D^2\left(\left\|u^{\epsilon}\right\|_{H^2(\Omega)}^2+\left\|u\right\|_{H^2(\Omega)}^2\right). \tag{5.74}$$

By Theorem 5.5,

$$\|u^\epsilon\|^2_{H^2(\Omega)} \leq \frac{8C_P^2+2}{\lambda^2}\|f\|^2_{L^2(\Omega)}. \tag{5.75}$$

Thus, combining inequalities (5.74) and (5.75), we have

$$\|e^\epsilon\|^2_{L^2(\Omega)} \leq \epsilon^2\, 8C_D^2\left(\frac{8C_P^2+2}{\lambda^2}\|f\|^2_{L^2(\Omega)} + \|u\|^2_{H^2(\Omega)}\right),$$

and the result follows. The proof is complete.

□

If $u^\epsilon \in H^3(\Omega)\cap V_g$ and $u \in H^3(\Omega)\cap V_g$, then we can expect to see a better rate of convergence in the H^1 norm as will be observed in the numerical experiments in Section 5.3. We can also derive convergence rates in the H^2 norm. However, the proof once again relies upon Conjecture 5.1.

Theorem 5.7. *Let $u^\epsilon \in H^3(\Omega)\cap V_g$ be the unique solution of (5.12). If $u \in H^3(\Omega)\cap V_g$ is the strong solution of (5.8), then*

$$\|\nabla(u-u^\epsilon)\|_{L^2(\Omega)} \leq \epsilon^{(1+\alpha)/2}\frac{c_3}{\lambda^{(1+\alpha)/2}}\left(\lambda^\alpha + c_4\right)^{1/2}\|u\|_{H^3(\Omega)}, \tag{5.76}$$

$$\left\|D^2(u-u^\epsilon)\right\|_{L^2(\Omega)} \leq \epsilon^{\alpha/2}\, c_5\left(1+\frac{c_6}{\lambda^\alpha}\right)^{1/2}\|u\|_{H^3(\Omega)}, \tag{5.77}$$

where c_1, c_2, c_3, c_4, c_5, and c_6 are positive constants which are independent of f, λ, u, and ϵ, and α is the dimension dependent constant from the trace inequality recorded in Section 5.2.1.

Proof. Pick $\epsilon > 0$ such that

$$\epsilon \leq \min\left\{\frac{2\lambda C_0^2}{3C_I^2C_P^2}, \left(\frac{\lambda}{6}\right)^{\frac{1}{1-\alpha}}\right\}.$$

Choose $\delta_A > 0$ such that

$$\delta_A \leq \min\left\{\frac{\epsilon}{4C_0^2\left(1+2C_A\right)}, \frac{\lambda}{3\left(C_P^2+2C_A+2C_AC_P^2\right)}\right\},$$

and fix M accordingly.

Let $v = -e^\epsilon \in V_0$ in (5.53). From Theorem 5.6, we can see

$$-\left(A : D^2 e^\epsilon, e^\epsilon\right)_\Omega = -\sum_{j=1}^{M} \Big((A - A_j) : D^2 e^\epsilon, e^\epsilon \Big)_{\Omega_j} - \sum_{j=1}^{M} \left(A_j : D^2 e^\epsilon, e^\epsilon\right)_{\Omega_j}, \tag{5.78}$$

with

$$\sum_{j=1}^{M} \Big((A - A_j) : D^2 e^\epsilon, e^\epsilon \Big)_{\Omega_j} \leq \frac{\delta_A}{2} \left\| D^2 e^\epsilon \right\|^2_{L^2(\Omega)} + \frac{\delta_A C_P^2}{2} \left\| \nabla e^\epsilon \right\|^2_{L^2(\Omega)}, \tag{5.79}$$

$$-\sum_{j=1}^{M} \left(A_j : D^2 e^\epsilon, e^\epsilon\right)_{\Omega_j} = -\sum_{j=1}^{M} \left(A_j \nabla e^\epsilon \cdot n_j, e^\epsilon\right)_{\partial\Omega_j} + \sum_{j=1}^{M} \left(A_j \nabla e^\epsilon, \nabla e^\epsilon\right)_{\Omega_j}, \tag{5.80}$$

$$\begin{aligned} \sum_{j=1}^{M} \left(A_j \nabla e^\epsilon \cdot n_j, e^\epsilon\right)_{\partial\Omega_j} &\leq \delta_A C_A \left\| e^\epsilon \right\|^2_{H^2(\Omega)} \\ &\leq \delta_A C_A \left\| D^2 e^\epsilon \right\|^2_{L^2(\Omega)} + \delta_A C_A \left(1 + C_P^2\right) \left\| \nabla e^\epsilon \right\|^2_{L^2(\Omega)}, \end{aligned} \tag{5.81}$$

and

$$\sum_{j=1}^{M} \left(A_j \nabla e^\epsilon, \nabla e^\epsilon\right)_{\Omega_j} \geq \lambda \left\| \nabla e^\epsilon \right\|^2_{L^2(\Omega)}, \tag{5.82}$$

where we have used Conjecture 5.1 to show (5.80).

We also have

$$\begin{aligned} -\epsilon \left(\Delta u^\epsilon, \Delta e^\epsilon\right)_\Omega &= -\epsilon \left(\Delta e^\epsilon, \Delta e^\epsilon\right)_\Omega - \epsilon \left(\Delta u, \Delta e^\epsilon\right)_\Omega \\ &= -\epsilon \left\| \Delta e^\epsilon \right\|^2_{L^2(\Omega)} - \epsilon \left(\Delta u, \Delta e^\epsilon\right)_\Omega \\ &= -\epsilon \left\| \Delta e^\epsilon \right\|^2_{L^2(\Omega)} - \epsilon \left(\Delta u, \nabla e^\epsilon \cdot n\right)_{\partial\Omega} + \epsilon \left(\nabla \Delta u, \nabla e^\epsilon\right)_\Omega \end{aligned} \tag{5.83}$$

with

$$\epsilon \left\|\Delta e^{\epsilon}\right\|^2_{L^2(\Omega)} \geq \frac{\epsilon}{C_0^2} \left\|D^2 e^{\epsilon}\right\|^2_{L^2(\Omega)}. \tag{5.84}$$

Observe,

$$\begin{aligned} -\epsilon \left(\Delta u, \nabla e^{\epsilon} \cdot n\right)_{\partial\Omega} &\leq \epsilon \left\|\Delta u\right\|_{L^2(\partial\Omega)} \left\|\nabla e^{\epsilon}\right\|_{L^2(\partial\Omega)} \\ &\leq \epsilon\, C_{\alpha} \left\|\Delta u\right\|_{L^2(\partial\Omega)} \left\|\nabla e^{\epsilon}\right\|^{\alpha}_{L^2(\Omega)} \left\|\nabla e^{\epsilon}\right\|^{1-\alpha}_{H^1(\Omega)} \\ &\leq \epsilon\, (1+\alpha)\, 2^{\frac{-2\alpha}{1+\alpha}} C_0^{\frac{2(1-\alpha)}{1+\alpha}} \left(C_{\alpha} \left\|\Delta u\right\|_{L^2(\partial\Omega)}\right)^{\frac{2}{1+\alpha}} \left\|\nabla e^{\epsilon}\right\|^{\frac{2\alpha}{1+\alpha}}_{L^2(\Omega)} \\ &\quad + \frac{\epsilon}{4C_0^2} \left\|\nabla e^{\epsilon}\right\|^2_{H^1(\Omega)} \\ &\leq \epsilon\, 2^{\frac{1-\alpha}{1+\alpha}} C_0^{\frac{2(1-\alpha)}{1+\alpha}} \left(C_{\alpha} C_T \left\|u\right\|_{H^3(\Omega)}\right)^{\frac{2}{1+\alpha}} \left\|\nabla e^{\epsilon}\right\|^{\frac{2\alpha}{1+\alpha}}_{L^2(\Omega)} \\ &\quad + \frac{\epsilon}{4C_0^2} \left\|\nabla e^{\epsilon}\right\|^2_{H^1(\Omega)}. \end{aligned} \tag{5.85}$$

Then,

$$\begin{aligned} &\epsilon\, 2^{\frac{1-\alpha}{1+\alpha}} C_0^{\frac{2(1-\alpha)}{1+\alpha}} \left(C_{\alpha} C_T \left\|u\right\|_{H^3(\Omega)}\right)^{\frac{2}{1+\alpha}} \left\|\nabla e^{\epsilon}\right\|^{\frac{2\alpha}{1+\alpha}}_{L^2(\Omega)} \\ &\quad \leq \frac{\lambda}{3} \left\|\nabla e^{\epsilon}\right\|^2_{L^2(\Omega)} + \frac{\epsilon^{1+\alpha} 3^{\alpha} 2^{1-\alpha} C_0^{2(1-\alpha)} C_{\alpha}^2 C_T^2 \left\|u\right\|^2_{H^3(\Omega)}}{\lambda^{\alpha}}, \end{aligned} \tag{5.86}$$

$$\begin{aligned} \frac{\epsilon}{4C_0^2} \left\|\nabla e^{\epsilon}\right\|^2_{H^1(\Omega)} &= \frac{\epsilon}{4C_0^2} \left(\left\|\nabla e^{\epsilon}\right\|^2_{L^2(\Omega)} + \left\|D^2 e^{\epsilon}\right\|^2_{L^2(\Omega)}\right) \\ &\leq \frac{\epsilon}{4C_0^2} \left(C_I \left\|e^{\epsilon}\right\|_{L^2(\Omega)} \left\|D^2 e^{\epsilon}\right\|_{L^2(\Omega)} + \left\|D^2 e^{\epsilon}\right\|^2_{L^2(\Omega)}\right) \\ &\leq \frac{\epsilon}{4C_0^2} \left(C_I C_P \left\|\nabla e^{\epsilon}\right\|_{L^2(\Omega)} \left\|D^2 e^{\epsilon}\right\|_{L^2(\Omega)} + \left\|D^2 e^{\epsilon}\right\|^2_{L^2(\Omega)}\right) \\ &\leq \frac{\epsilon\, C_I C_P}{C_0^2} \left\|\nabla e^{\epsilon}\right\|_{L^2(\Omega)} \left\|D^2 e^{\epsilon}\right\|_{L^2(\Omega)} + \frac{\epsilon}{4C_0^2} \left\|D^2 e^{\epsilon}\right\|^2_{L^2(\Omega)}, \end{aligned} \tag{5.87}$$

and

$$\frac{\epsilon\, C_I C_P}{C_0^2} \left\|\nabla e^\epsilon\right\|_{L^2(\Omega)} \left\|D^2 e^\epsilon\right\|_{L^2(\Omega)} \leq \frac{\epsilon\, C_I^2 C_P^2}{8C_0^2} \left\|\nabla e^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\epsilon}{8C_0^2} \left\|D^2 e^\epsilon\right\|_{L^2(\Omega)}^2. \tag{5.88}$$

Lastly,

$$\begin{aligned} \epsilon \left(\nabla \Delta u, \nabla e^\epsilon\right)_\Omega &\leq \epsilon \left\|\nabla \Delta u\right\|_{L^2(\Omega)} \left\|\nabla e^\epsilon\right\|_{L^2(\Omega)} \\ &\leq \frac{\epsilon^{1-\alpha}}{2} \left\|\nabla e^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\epsilon^{1+\alpha}}{2} \left\|D^3 u\right\|_{L^2(\Omega)}^2 \\ &\leq \frac{\epsilon^{1-\alpha}}{2} \left\|\nabla e^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\epsilon^{1+\alpha}}{2} \left\|u\right\|_{H^3(\Omega)}^2. \end{aligned} \tag{5.89}$$

Thus, combining (5.78) - (5.89) and substituting into (5.53), we have

$$\begin{aligned} &\frac{5\epsilon}{8C_0^2} \left\|D^2 e^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{2\lambda}{3} \left\|\nabla e^\epsilon\right\|_{L^2(\Omega)}^2 \\ &\quad\leq \frac{\delta_A \left(1 + 2C_A\right)}{2} \left\|D^2 e^\epsilon\right\|_{L^2(\Omega)}^2 \\ &\qquad + \left[\frac{\delta_A \left(C_P^2 + 2C_A + 2C_A C_P^2\right)}{2} + \frac{\epsilon\, C_I^2 C_P^2}{8C_0^2} + \frac{\epsilon^{1-\alpha}}{2}\right] \left\|\nabla e^\epsilon\right\|_{L^2(\Omega)}^2 \\ &\qquad + \frac{\lambda^\alpha + 3^\alpha 2^{2-\alpha} C_0^{2(1-\alpha)} C_\alpha^2 C_T^2}{2\lambda^\alpha} \left\|u\right\|_{H^3(\Omega)}^2 \epsilon^{1+\alpha}. \end{aligned} \tag{5.90}$$

Using the restrictions on ϵ and δ_A, we have

$$\frac{\epsilon}{2C_0^2} \left\|D^2 e^\epsilon\right\|_{L^2(\Omega)}^2 + \frac{\lambda}{3} \left\|\nabla e^\epsilon\right\|_{L^2(\Omega)}^2 \leq \frac{\lambda^\alpha + 3^\alpha 2^{2-\alpha} C_0^{2(1-\alpha)} C_\alpha^2 C_T^2}{2\lambda^\alpha} \left\|u\right\|_{H^3(\Omega)}^2 \epsilon^{1+\alpha}. \tag{5.91}$$

Therefore,

$$\left\|D^2 e^\epsilon\right\|_{L^2(\Omega)}^2 \leq \epsilon^\alpha C_0^2 \left(1 + \frac{3^\alpha 2^{2-\alpha} C_0^{2(1-\alpha)} C_\alpha^2 C_T^2}{\lambda^\alpha}\right) \left\|u\right\|_{H^3(\Omega)}^2$$

and

$$\left\|\nabla e^\epsilon\right\|_{L^2(\Omega)}^2 \leq \epsilon^{1+\alpha} \frac{3}{2\,\lambda^{1+\alpha}} \left(\lambda^\alpha + 3^\alpha 2^{2-\alpha} C_0^{2(1-\alpha)} C_\alpha^2 C_T^2\right) \left\|u\right\|_{H^3(\Omega)}^2,$$

and the bounds for $\|D^2 e^\epsilon\|_{L^2(\Omega)}$ and $\|\nabla e^\epsilon\|_{L^2(\Omega)}$ follow. The proof is complete.

□

Remark 5.2. *If $n = 2$, then $\alpha = 1/2$. Thus, for $n = 2$, u^ϵ is an $\mathcal{O}\left(\epsilon^{3/4}\right)$ approximation of u in the H^1 norm if $u \in H^3(\Omega)$.*

5.2.5 Benefits of the Methodology

We have shown that problem (5.12) has a unique solution for every $\epsilon > 0$. Additionally, if u is the unique strong solution of (5.8), then the family of solutions u^ϵ converges to u in $H^1(\Omega)$ as $\epsilon \to 0$. Thus, the linear fourth order PDE problem (5.12) approximates the linear second order PDE problem (5.8) that has non-divergence form.

In the previous section we have actually quantified the rates of convergence with respect to ϵ. Suppose u is the solution to (5.8), u^ϵ is the solution to (5.12), and u_h^ϵ is a numerical approximation of u^ϵ. Then, the error involved in approximating u by u_h^ϵ can be split into two components, the PDE approximation error characterized by ϵ and the numerical approximation error characterized by h. Thus, we have

$$\|u - u_h^\epsilon\| \leq \|u - u^\epsilon\| + \|u^\epsilon - u_h^\epsilon\| \leq C_1\, \epsilon^r + C_2(\epsilon)\, h^s,$$

where the rate r is determined by the choice of the norm and the rate s is determined by the numerical method. We note that $C_2(\epsilon)$ may blow-up in ϵ depending upon the choice of the norm $\|\cdot\|$. Letting $\epsilon = h^{s/r}$ and assuming C_2 is independent of ϵ, we have u_h^ϵ is an s-order approximation of u. Thus, we have developed an analytic framework for the error analysis of a given numerical scheme for approximating (5.12).

Lastly, we remark that problem (5.12) is composed of a biharmonic type equation that can be approximated by several well-studied numerical methods including conforming finite element methods such as the Hermite element in one-dimension or the Argyris element in two-dimensions, nonconforming finite element methods such

as the Morley element in two-dimensions, mixed finite element methods, spectral methods, discontinuous Galerkin methods, etc. By applying the vanishing moment method, we are able to transform a second order problem in strong form to a fourth order problem in weak form that is readily approximated by many Galerkin-based numerical methods, while the original problem is not directly accessible by such methods as will be demonstrated in Examples 5.3 and 5.4 below.

5.3 Numerical Experiments

In this section, we provide a series of numerical tests that demonstrate the effectiveness of the methodologies presented in this chapter. First we will test the conclusions in Section 5.2 regarding the use of the vanishing moment method for approximating the strong solution of second order linear elliptic PDEs of non-divergence form in two-dimensions. Then, in the next section, we will test Algorithm 5.1 for approximating stationary HJB problems in one-dimension. All of the numerical tests in this section were performed using COMSOL.

5.3.1 Linear Elliptic Equations of Non-Divergence Form

We now perform a series of numerical tests to support the conclusions of Theorems 5.6 and 5.7. To this end, we perform the following experiment using the fifth order Argyris finite element space (see [9]) to approximate (5.12). Let $\mathcal{T}_h$ be a fine mesh for the domain Ω. We fix a continuous, positive definite matrix A that is neither symmetric nor differentiable. Then, we approximate problem (5.12) with known solutions in H^2 and H^3 for varying values of ϵ. The tests are based on the ansatz that with a fine enough mesh, the approximation error will be dominated by the PDE approximation resulting from using the vanishing moment method and that the error due to the

numerical scheme will be negligible in comparison. We will fix

$$A(x,y) = \begin{bmatrix} (2x-y)^{1/3} + 4e^{2-x} & \sin(10xy) \\ -(x+2)^{1/2} & |y-2x|^{1/4} + 3 \end{bmatrix}$$

for the first three tests, and then choose A to be non-uniformly elliptic for the fourth test as a means to test the impact of the method for "harder" problems not considered in the purely theoretical sections above. We will observe that the error is maximized along sets where the solution is not as regular and along the boundary due to the auxiliary boundary condition (5.10c). The measured error does not appear to correspond to sets where the coefficient matrix A is not as well-behaved.

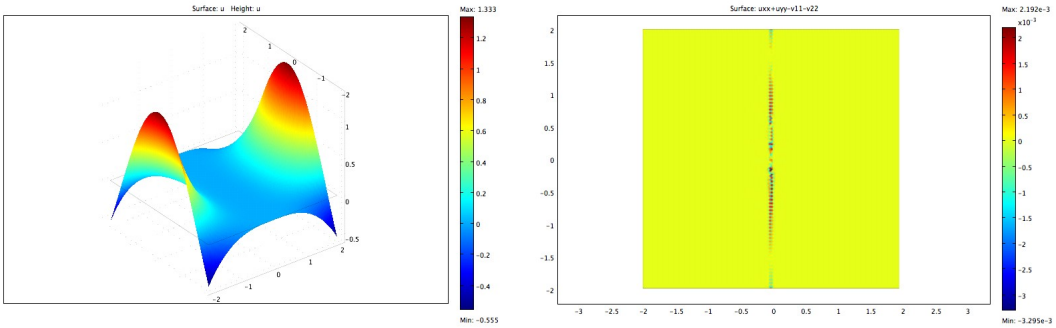
Example 5.1. *Consider* (5.12) *with* $\Omega = (-2,2) \times (-2,2)$ *and solution* $u(x,y) = \frac{1}{6}|x|^3 \cos(y) \in H^3(\Omega)$*, for* f *and* g *chosen accordingly.*

The given test problem has a solution in $H^3(\Omega)$, where the third-order partial derivative with respect to x is discontinuous along the line $x = 0$. Approximating for various values of ϵ, we observe optimal convergence rates as found in Table 5.1. We can also see from Figure 5.1 that the error is largest along the line $x = 0$, as expected.

Table 5.1: Rates of convergence for Example 5.1 using the vanishing moment method.

ϵ	$\|u^\epsilon - u\|_{L^2(\Omega)}$	Order	$\|\nabla(u^\epsilon - u)\|_{L^2(\Omega)}$	Order	$\|\Delta(u^\epsilon - u)\|_{L^2(\Omega)}$	Order
4e-2	9.44e-3		2.25e-2		2.55e-1	
2e-2	4.89e-3	0.95	1.32e-2	0.76	2.15e-1	0.24
1e-2	2.50e-3	0.96	7.84e-3	0.76	1.82e-1	0.24
5e-3	1.27e-3	0.98	4.64e-3	0.76	1.54e-1	0.25
2.5e-7	1.65e-7		2.09e-5		3.80e-3	

Example 5.2. *Consider* (5.12) *with* $\Omega = (-2,2) \times (-2,2)$ *and solution* $u(x,y) = \frac{1}{2}x|x|\cos(y) \in H^2(\Omega)$*, for* f *and* g *chosen accordingly.*



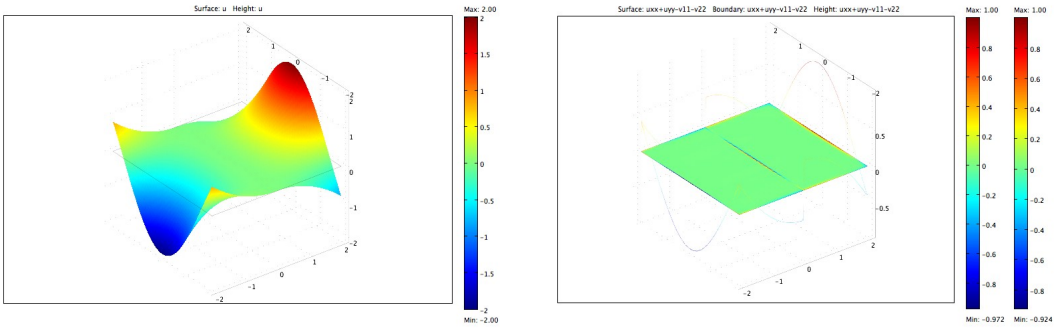
(a) u_h^ϵ. **(b)** $\Delta\left(u - u_h^\epsilon\right)$.

Figure 5.1: Computed solution and corresponding error for Example 5.1 using the vanishing moment method.

The given test problem has a solution in $H^2(\Omega)$, where the second-order derivative is discontinuous along the line $x = 0$. Approximating for various values of ϵ, we observe slightly sub-optimal convergence rates for the L^2 norm and convergence rates between the theoretical rates of Theorems 5.6 and 5.7 for the H^1 norm in Table 5.2. We also have a rate of convergence in the H^2 norm that is consistent with assuming $H^3(\Omega)$ regularity for the solution. Figure 5.2 shows that the error is once again largest along the line $x = 0$. The figure also shows the boundary layer due to the high-order auxiliary boundary condition.

Table 5.2: Rates of convergence for Example 5.2 using the vanishing moment method.

ϵ	$\left\|u^\epsilon - u\right\|_{L^2(\Omega)}$	Order	$\left\|\nabla\left(u^\epsilon - u\right)\right\|_{L^2(\Omega)}$	Order	$\left\|\Delta\left(u^\epsilon - u\right)\right\|_{L^2(\Omega)}$	Order
4e-2	4.30e-3		2.26e-2		3.99e-1	
2e-2	2.30e-3	0.90	1.42e-2	0.67	3.38e-1	0.24
1e-2	1.22e-3	0.91	8.79e-3	0.69	2.86e-1	0.24
5e-7	2.25e-4		9.09e-4		1.07e-1	



(a) u_h^ϵ. **(b)** $\Delta\left(u - u_h^\epsilon\right)$.

Figure 5.2: Computed solution and corresponding error for Example 5.2 using the vanishing moment method.

Example 5.3. *Consider* (5.12) *with* $\Omega = B_2(0)$*, the ball of radius 2 centered at the origin, and solution* $u(x, y) = (x - y)^{8/3} \in H^2(\Omega)$*, for* f *and* g *chosen accordingly.*

Observe, the solution is once again in $H^2(\Omega)$, where the second order derivatives have a cusp along the line $x = y$. Also, the domain is a disc. Approximating for various values of ϵ, we observe the convergence rates of Theorem 5.7, where H^3 regularity is assumed, in Table 5.3. We can also see the finite element method does not converge to u when using $\epsilon = 0$, which verifies the fact that the Argyris finite element method does not work for second order linear problems of non-divergence form, even when approximating an H^2 solution. The plot of an approximation can be found in Figure 5.3.

Example 5.4. *Consider* (5.12) *with* $\Omega = (-2, 2) \times (-2, 2)$*,*

$$A(x, y) = \frac{16}{9} \begin{bmatrix} x^{2/3} & -x^{1/3}y^{1/3} \\ -x^{1/3}y^{1/3} & y^{2/3} \end{bmatrix},$$

and solution $u(x, y) = x^{4/3} - y^{4/3} \in H^1(\Omega)$*, for* f *and* g *chosen accordingly.*

Table 5.3: Rates of convergence for Example 5.3 using the vanishing moment method.

ϵ	$\|u^\epsilon - u\|_{L^2(\Omega)}$	Order	$\|\nabla(u^\epsilon - u)\|_{L^2(\Omega)}$	Order	$\|\Delta(u^\epsilon - u)\|_{L^2(\Omega)}$	Order
5e-3	1.98e-2		1.22e-1		4.27	
2.5e-3	1.01e-2	0.98	7.32e-2	0.73	3.59	0.25
1.25e-3	5.08e-3	0.98	4.40e-2	0.73	3.03	0.24
0	5.40e3		1.83e6		4.99e8	

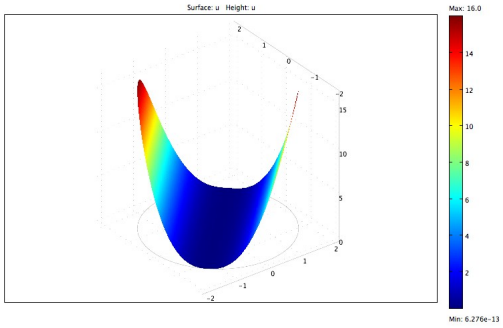
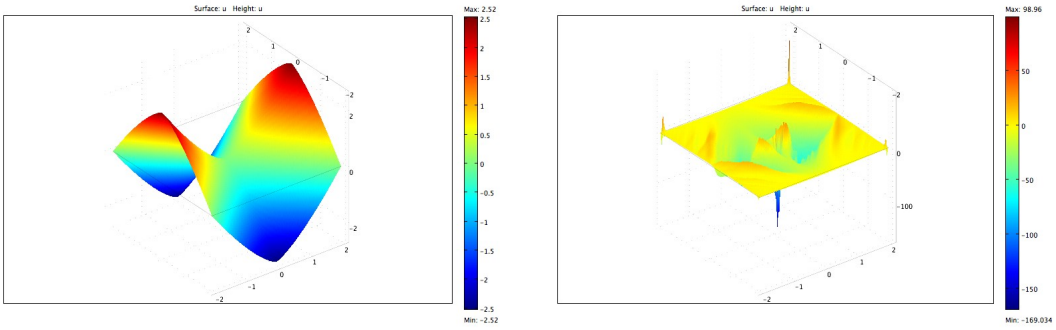


Figure 5.3: Computed solution for Example 5.3 using the vanishing moment method.

For the above example, we have $u \notin H^2(\Omega)$ and A is not uniformly elliptic. Thus, u is only a viscosity solution, not a strong solution. We see in Table 5.4 that the vanishing moment method appears to be working, although with unknown deteriorated rates of convergence. From Figure 5.4, we see that the finite element method with Argyris finite element space does not work for the given example. Thus, we can see that the vanishing moment method has strong potential for approximating more general second order problems that are understood in the viscosity solution framework.

Table 5.4: Rates of convergence for Example 5.4 using the vanishing moment method.

ϵ	$\|u^\epsilon - u\|_{L^2(\Omega)}$	Order	$\|\nabla(u^\epsilon - u)\|_{L^2(\Omega)}$	Order
2e-4	6.89e-2		3.94e-1	
1e-4	6.10e-2	0.18	3.53e-1	0.16
5e-5	5.46e-2	0.16	3.15e-1	0.17



(a) u_h^ϵ. **(b)** u_h^ϵ for $\epsilon = 0$.

Figure 5.4: Computed solutions for Example 5.4 using the vanishing moment method.

5.3.2 Static Hamilton-Jacobi-Bellman Equations in One-Dimension

We now perform a series of tests for approximating the stationary HJB problem (5.3) using Algorithm 5.1. To this end, we use COMSOL to implement the finite element method with the third-order Hermite finite element space, which is a C^1 element in one-dimension. We also use Livelink as a means for solving the optimization step of Algorithm 5.1 using Matlab's built-in functions *fminsearch* and *fminbnd*. For each test problem, we fix our approximation values ϵ and h with an initial guess for θ_h, and then we record the results of Algorithm 5.1 over several iterations. We also record

the L^1 norm of $L_{\theta_h^*} u_h^*$ using the final iteration, which appears to be a good stopping parameter for Algorithm 5.1.

Example 5.5. *Consider the problem*

$$\inf_{\theta(x)\in\mathbb{R}} \left\{-u_{xx} + \theta^2 u\right\} = 0, \qquad -1 < x < 2,$$
$$u(-1) = u(2) = 1$$

with solution $u^*(x) = 1$ *corresponding to* $\theta^*(x) = 0$.

The first example has constant solutions, and we can see from Table 5.5 and Figure 5.5 that the solution is captured in one iteration. We also have $\left\|L_{\theta_h^*} u_h^*\right\|_{L^1(\Omega)} =$ 6.501361e-10.

Table 5.5: Performance of Algorithm 5.1 for Example 5.5 with $\epsilon =$ 1.0e-8, $h = 0.005$, and initial guess $\theta_h^{(0)}(x) = 1$.

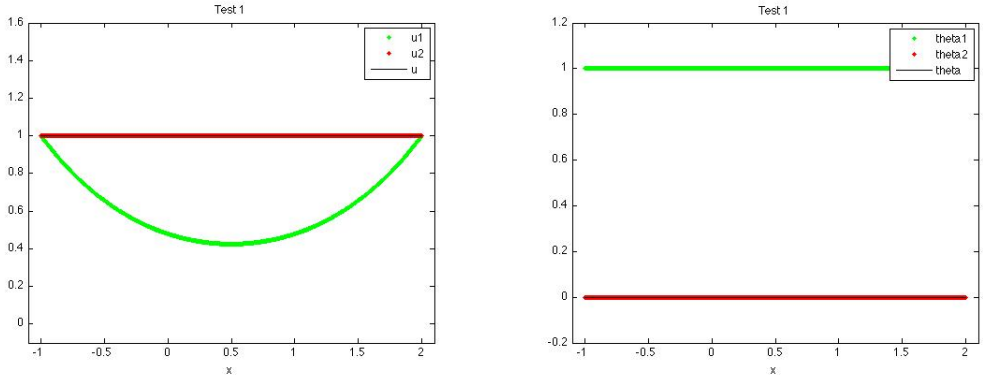
iteration	$\|u_h - u\|_{L^2(\Omega)}$	$\|\theta_h - \theta\|_{L^2(\Omega)}$
1	0.7454	1.7321
2	2.7024e-11	9.4931e-17
3	5.6866e-11	3.2802e-17

Example 5.6. *Consider the problem*

$$\inf_{\theta(x)\in\mathbb{R}^2} \left\{-u_{xx} + \theta_1^2 u + (\theta_2 - 2)^2\right\} = 0, \qquad -1 < x < 2,$$
$$u(-1) = 1 + \theta_2(-1)^2, \qquad u(2) = 1 + \theta_2(2)^2$$

with solution $u^*(x) = 5$ *corresponding to* $\theta_1^*(x) = 0$ *and* $\theta_2^*(x) = 2$.

The second example again has constant solutions, but now the optimization is performed over a two-dimensional space. From Table 5.6 and Figure 5.6, we again



(a) u_h^ϵ. **(b)** θ_h^ϵ.

Figure 5.5: Computed solutions for Example 5.5 using Algorithm 5.1 with ϵ = 1.0e-8, $h = 0.005$, and initial guess $\theta_h^{(0)}(x) = 1$.

see that we capture the exact solution in only one iteration with $\left\|L_{\theta_h^*} u_h^*\right\|_{L^1(\Omega)}$ = 4.927490e-9.

Table 5.6: Performance of Algorithm 5.1 for Example 5.6 with ϵ = 1.0e-8, $h = 0.005$, and initial guess $\theta_1^{(0)}(x) = 1$, $\theta_2^{(0)}(x) = 1$.

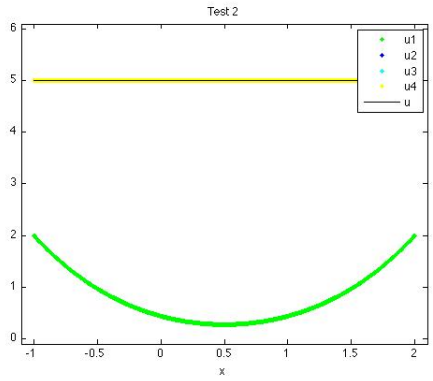
iteration	$\|u_h - u\|_{L^2(\Omega)}$	$\|\theta_h - \theta\|_{L^2(\Omega)}$
1	7.3086	6.0000
2	3.3620e-09	1.2301e-09
3	3.9834e-09	1.2311e-09
4	3.3620e-09	1.2311e-09

Example 5.7. *Consider the problem*

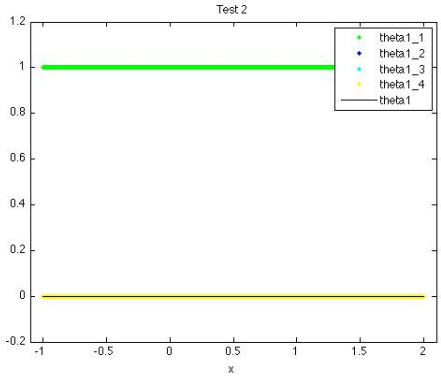
$$\inf_{\theta(x)\in\mathbb{R}} \left\{-\theta u_{xx} + \theta^2 u - x^{-2}\right\} = 0, \qquad 2 < x < 4,$$

$$u(2) = 4, \qquad u(4) = 16$$

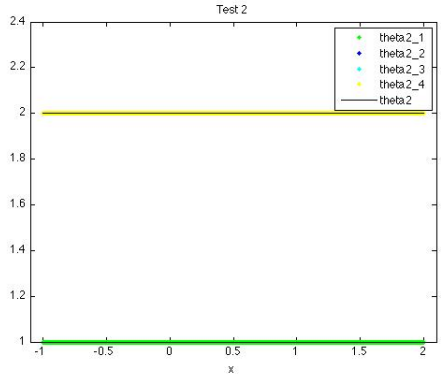
with solution $u^*(x) = x^2$ *corresponding to* $\theta^*(x) = x^{-2}$.



(a) u_h^ϵ.



(b) θ_1^ϵ.



(c) θ_2^ϵ.

Figure 5.6: Computed solutions for Example 5.6 using Algorithm 5.1 with $\epsilon =$ 1.0e-8, $h = 0.005$, and initial guess $\theta_1^{(0)}(x) = 1$, $\theta_2^{(0)}(x) = 1$.

The third example involves an unbounded optimization space with exact solutions that are not constant. From Table 5.7 and Figure 5.7, we can see that the algorithm appears to be converging with the use of a few more iterations. We have $\left\| L_{\theta_h^*} u_h^* \right\|_{L^1(\Omega)} =$ 2.690732e-5 for the ninth iteration.

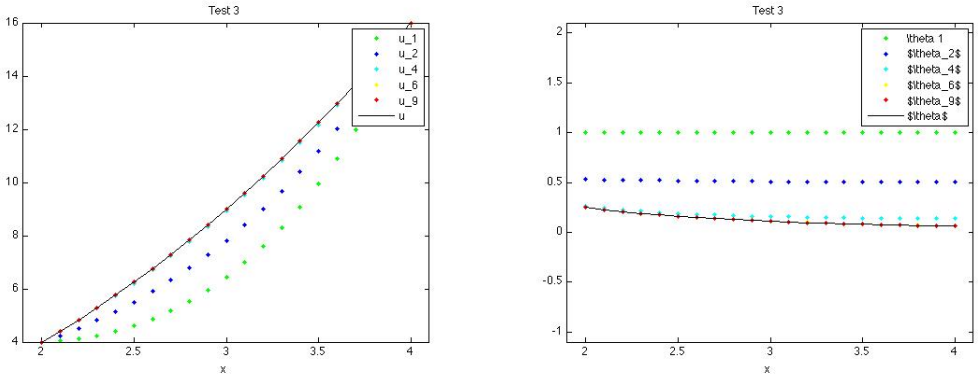
Example 5.8. *Consider the problem*

$$\inf_{-1 \le \theta(x) \le 1} \left\{ |x-1| u_{xx} + \theta u_x - 3|x-1|^2 \right\} = 0, \qquad 0 < x < 2.4,$$

$$u(0) = 1, \qquad u(2.4) = 2.744$$

Table 5.7: Performance of Algorithm 5.1 for Example 5.7 with ϵ = 1.0e-8, $h = 0.1$, and initial guess $\theta_h^{(0)}(x) = 1$.

iteration	$\|u_h - u\|_{L^2(\Omega)}$	$\|\theta_h - \theta\|_{L^2(\Omega)}$
1	2.710562e+00	1.239540e+00
2	1.256370e+00	5.497131e-01
4	8.165944e-02	6.880879e-02
6	4.097456e-05	1.326518e-03
9	5.485387e-07	2.587541e-04



(a) u_h^ϵ. **(b)** θ_h^ϵ.

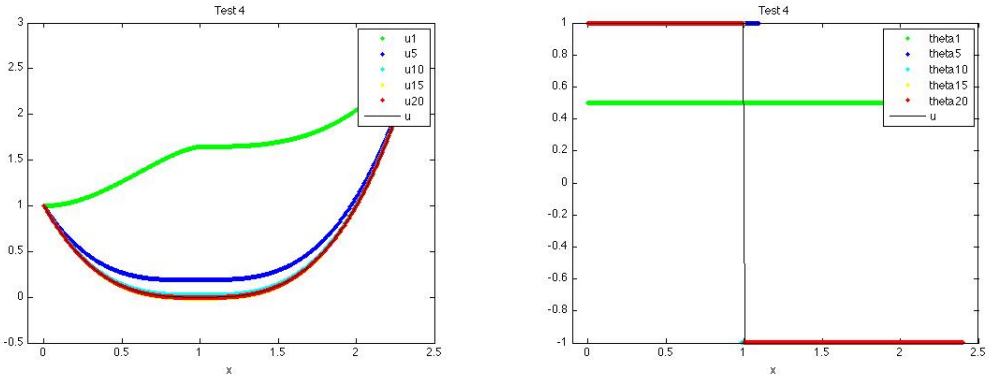
Figure 5.7: Computed solutions for Example 5.7 using Algorithm 5.1 with ϵ = 1.0e-8, $h = 0.1$, and initial guess $\theta_h^{(0)}(x) = 1$.

with solution $u^*(x) = |x-1|^3$ *corresponding to* $\theta^*(x) = -\text{sign}(x)$.

The last example features a bounded optimization space where the optimal values are along the boundary of Θ. The results are recorded in Table 5.8 and Figure 5.8. Now the algorithm requires many more iterations to converge. After 20 iterations we have $\left\|L_{\theta_h^*} u_h^*\right\|_{L^1(\Omega)} = 1.505992\text{e-}5$.

Table 5.8: Performance of Algorithm 5.1 for Example 5.8 with ϵ = 1.0e-8, $h = 0.005$, and initial guess $\theta_h^{(0)}(x) = 0$.

iteration	$\|u_h - u\|_{L^2(\Omega)}$	$\|\theta_h - \theta\|_{L^2(\Omega)}$
1	1.936128e+00	1.843909e+00
5	2.161311e-01	6.324401e-01
10	2.554534e-02	2.943850e-01
15	1.173412e-02	8.164867e-02
20	4.122284e-03	8.164867e-02



(a) u_h^ϵ. **(b)** θ_h^ϵ.

Figure 5.8: Computed solutions for Example 5.8 using Algorithm 5.1 with ϵ = 1.0e-8, $h = 0.005$, and initial guess $\theta_h^{(0)}(x) = 0$.

Chapter 6

Summary and Future Directions

The research presented in this dissertation primarily focusses on the development of direct numerical methods for approximating the viscosity solutions of fully nonlinear second order elliptic and parabolic PDEs. In particular, approximation frameworks for fully nonlinear second order elliptic PDE problems with Dirichlet boundary data are developed, and then the frameworks are extended to the corresponding parabolic equations using the method of lines for the time discretization. The methods and results presented in this dissertation are expected to have significant impacts in the scientific community due to the theoretical contributions and positive numerical tests regarding the Monge-Ampère equation and the Hamilton-Jacobi-Bellman equation, two prototypical fully nonlinear second order PDEs that arise in many application problems.

In this dissertation, a complete FD convergence framework is established for directly approximating continuous viscosity solutions of one-dimensional fully nonlinear second order elliptic PDEs that satisfy the comparison principle, and the framework is formally extended for PDEs in high dimensions. Key ideas from the FD framework include numerical operators, numerical moments, and generalized monotonicity. Borrowing the ideas from the FD framework, nonstandard LDG and IPDG methods are also developed for directly approximating fully nonlinear second

order PDEs. The LDG methods are shown to be natural high order analogues of their FD counterparts. Lastly, a vanishing moment methodology is developed for indirectly approximating Hamilton-Jacobi-Bellman equations.

As expected, the pursued research has led to more questions than could be answered during the time frame for completing the dissertation. The presented theory in Chapter 2 is only targeted at showing convergence. Analytic techniques for deriving convergence rates still remain to be developed. Also, only the Lax-Friedrichs-like numerical operator is analyzed with regards to admissibility and stability. The LDG methods of Chapter 3, as well as the LDG methods of Yan and Osher for Hamilton-Jacobi equations, are only known to converge when using a uniform Cartesian partition with piecewise constant basis functions due to the equivalence with convergent FD methods. Thus, general techniques must be developed for analyzing the convergence of the LDG methods for both first and second order problems. The potential for the adaptivity of the formulated DG methods has yet to be investigated. The numerical tests in Chapters 3 and 4 also demonstrate the need to further develop and analyze nonlinear solvers for the proposed LDG and IPDG methods. In many ways, the discretization of fully nonlinear PDEs and the choice/design of nonlinear solvers are best thought of as two sides of the same coin. Numerical issues linked with numerical artifacts may require efforts during both the discretization process and the design of the nonlinear solver, two areas of numerical PDEs that are often treated independently.

The remainder of the chapter focusses on three of the applications and questions that were directly referenced in the dissertation. We first explore some analytic issues that arise when expanding the convergence analysis of the FD framework to include the formulation for PDEs in high dimensions. Next, we record a few results regarding the discontinuous Galerkin finite element (DGFE) differential calculus and the symmetric dual-wind discontinuous Galerkin (DWDG) method that were both partly inspired by the research effort for this dissertation. Last, we discuss some future directions concerning the topics discussed in Chapter 5.

6.1 Convergence of the High-Dimensional Finite Difference Framework

We now discuss the analytic issues concerning the FD framework proposed in Section 2.5. In particular, we consider the FD method represented by (2.68). There are two obstacles that prevent a natural generalization of the one-dimensional convergence proof given in Section 2.3.2. The first involves comparing the two discrete Hessian operators $\widetilde{D}_h^2$ and $\overline{D}_h^2$ in conjunction with the g-monotonicity and consistency assumptions. The second involves showing a discrete Hessian approximation is negative semidefinite when acting on an upper semi-continuous function at a relative maximum. Based on the counter-examples presented in Sections 6.1.1 and 6.1.2, we will see that the one-dimensional convergence proof does not provide a sufficient template that can be used in a trivial way to prove convergence for the high-dimensional FD framework. However, the relationship of the high-dimensional FD framework with the vanishing moment methodology and the positive numerical tests provided in Chapter 3 with $r = 0$ yield evidence that the new FD methodology may be convergent even though a proof is not currently known. Furthermore, we will see in the following that the high-dimensional FD framework is convergent to the unique viscosity solution u when $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ and the piecewise constant extension functions defined by (2.70) converge to a function in $C^2(\Omega) \cap C^0(\overline{\Omega})$. Thus, more research for less regular solutions needs to be performed.

6.1.1 Comparing Discrete Hessians

A key component of the one-dimensional FD convergence proof given in Section 2.3.2 is (2.37), which yields the relationship

$$\lim_{k\to\infty} \tilde{\delta}^2_{x,h_k} u_{h_k}(\xi_k) \geq \lim_{k\to\infty} \delta^2_{x,h_k} u_{h_k}(\xi_k).$$

Thus, a natural generalization for (2.68) would be

$$\lim_{k\to\infty} \widetilde{D}^2_{x,h_k} u_{h_k}(\xi_k) \geq \lim_{k\to\infty} \overline{D}^2_{x,h_k} u_{h_k}(\xi_k),$$

where $A \geq B$ implies $A - B$ is positive semidefinite for A, B symmetric matrices. However, we will see in the following Lemma that such an inequality does not exist when $\overline{u} := \limsup u_{h_k}$ has low regularity. Furthermore, we will see in the following lemma that we could also use the convention that $A \geq B$ implies $[A - B]_{i,j} \geq 0$ for all $i, j = 1, 2, \ldots, d$, i.e., $A \geq B$ component-wise, without a change in the observations. Under this new idea of an ordering for matrices, g-monotonicity would become a component-wise monotonicity requirement.

Lemma 6.1. *Let $v : \Omega \to \mathbb{R}$ have a relative maximum at $(x_0, y_0) \in \Omega \subset \mathbb{R}^2$. Then the upper semi-continuity of v is not sufficient to guarantee there exists sequences $\{h_{x_k}\}_{k\geq 1}$, $\{h_{y_k}\}_{k\geq 1}$ such that $h_{x_k}, h_{y_k} \to 0$ and*

$$\lim_{k\to\infty} \widetilde{D}^2_{x,h_k} v(x_0, y_0) \geq \lim_{k\to\infty} \overline{D}^2_{x,h_k} v(x_0, y_0) \tag{6.1}$$

or

$$\lim_{k\to\infty} \widetilde{D}^2_{x,h_k} v(x_0, y_0) < \lim_{k\to\infty} \overline{D}^2_{x,h_k} v(x_0, y_0), \tag{6.2}$$

where the ordering is understood either using the natural ordering for symmetric matrices or component-wise.

Proof. Let $(x_0, y_0) = (0, 0)$ and $\Omega = B_1(0)$, the unit ball centered at the origin.

Define $v_1 : \Omega \to \mathbb{R}$ by

$$v_1(x) = \begin{cases} -x^2 - y^2, & \text{if } xy \geq 0, \\ -x^2 - y^2 - 1, & \text{otherwise.} \end{cases}$$

Then v_1 is upper semi-continuous on Ω. Observe,

$$\widetilde{D}_h^2 v_1(0,0) = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$$

and

$$\overline{D}_h^2 v_1(0,0) = \begin{bmatrix} -2 & -\frac{h_x}{h_y} - \frac{h_y}{h_x} + 2\frac{h_x^2+h_y^2}{h_x h_y} + \frac{1}{h_x h_y} \\ -\frac{h_x}{h_y} - \frac{h_y}{h_x} + 2\frac{h_x^2+h_y^2}{h_x h_y} + \frac{1}{h_x h_y} & -2 \end{bmatrix}.$$

Thus,

$$\widetilde{D}_h^2 v_1(0,0) - \overline{D}_h^2 v_1(0,0) = \begin{bmatrix} 0 & \frac{h_x}{h_y} + \frac{h_y}{h_x} - 2\frac{h_x^2+h_y^2}{h_x h_y} - \frac{1}{h_x h_y} \\ \frac{h_x}{h_y} + \frac{h_y}{h_x} - 2\frac{h_x^2+h_y^2}{h_x h_y} - \frac{1}{h_x h_y} & 0 \end{bmatrix},$$

and we have $\widetilde{D}_h^2 v_1(0,0) - \overline{D}_h^2 v_1(0,0)$ is neither positive semidefinite nor negative semidefinite for all $h > 0$.

Define $v_2 : \Omega \to \mathbb{R}$ by

$$v_2(x) = \begin{cases} -x^2, & \text{if } y = 0, \\ -|x| - 1, & \text{otherwise} \end{cases}$$

and $v_3 : \Omega \to \mathbb{R}$ by

$$v_3(x) = \begin{cases} -|x|, & \text{if } y = 0, \\ -1, & \text{otherwise.} \end{cases}$$

Then v_2 and v_3 are upper semi-continuous on Ω. A simple computation reveals

$$\left[\widetilde{D}_h^2 v_i(0,0) - \overline{D}_h^2 v_i(0,0)\right]_{1,2} = \frac{h_x}{2h_y}\left(\delta_{x,h_x}^2 v_i(0,h_y) + \delta_{x,h_x}^2 v_i(0,-h_y) - 2\delta_{x,h_x}^2 v_i(0,0)\right)$$

for $i = 2, 3$. Thus,

$$\left[\widetilde{D}_h^2 v_2(0,0) - \overline{D}_h^2 v_2(0,0)\right]_{1,2} = \frac{2h_x - 2}{h_y} \to -\infty$$

and

$$\left[\widetilde{D}_h^2 v_3(0,0) - \overline{D}_h^2 v_3(0,0)\right]_{1,2} = \frac{2}{h_y} \to \infty.$$

Hence, we do not have a consistent way to compare the off-diagonal elements of $\widetilde{D}_h^2 v$ and $\overline{D}_h^2 v$ for v upper semi-continuous, and it follows that neither (6.1) nor (6.2) hold in general when using a component-wise ordering. The proof is complete. □

We could avoid the issue of comparing off-diagonal elements of the Hessian approximations by using another discrete Hessian approximation defined by (2.12), i.e.,

$$\left[\widehat{D}_h^2\right]_{i,j} = \begin{cases} \frac{\delta^+_{x_i,h_i}\delta^+_{x_i,h_i} + \delta^-_{x_i,h_i}\delta^-_{x_i,h_i}}{2}, & \text{if } i = j, \\ \delta^2_{x_i,h_i;x_j,h_j}, & \text{otherwise,} \end{cases} \tag{6.3}$$

paired with the alternative FD method: Find a grid function U such that

$$\widehat{G}\left(\widehat{D}_h^2 U_\alpha, D_h^2 U_\alpha, \nabla_h^+ U_\alpha, \nabla_h^- U_\alpha, U_\alpha, x_\alpha\right) = 0 \tag{6.4}$$

for all $\alpha \in \mathbb{N}_J^d$ with $\alpha_i \in \{2, 3, \ldots, J_i - 1\}$ for all $i = 1, 2, \ldots, d$, where D_h^2 is defined in (2.10) and $\widehat{G}$ is a consistent, g-monotone numerical operator.

The only difference in (2.68) and (6.4) is how the off-diagonal elements of the discrete Hessians are computed. In (2.68), the off-diagonal elements of the discrete Hessians for the two arguments, $\widetilde{D}_h^2 U_\alpha$ and $\overline{D}_h^2 U_\alpha$, are not the same. A benefit of the off-diagonal terms not being equivalent is the ability to form true numerical moments that approximate the scaled biharmonic operator, as seen in Remark 2.5. Then, for a Lax-Friedrichs-like numerical operator, FD method (2.68) is a direct realization of the vanishing moment method. Conversely, in (6.4), the off-diagonal elements of the two discrete Hessian parameters are the same. Thus, we immediately have the

ability to select sequences $\{h_k\}_{k=1}^{\infty}$, $\{\xi_k\}_{k=1}^{\infty}$, and $\{\epsilon_k\}_{k=1}^{\infty}$ such that $\widehat{D}^2_{h_k} u_{h_k}(\xi_k) \geq D^2_{h_k} u_{h_k}(\xi_k) + \epsilon_k I$ and $\epsilon_k \to 0$ as in the one-dimensional convergence proof. However, using (6.4), we can no longer form a vanishing biharmonic approximation when using a Lax-Friedrichs-like numerical operator. Instead, the numerical moment would approximate the scaled fourth-order operator $\sum_{i=1}^{d} h_i^2 \frac{\partial^4}{\partial x_i^4}$. While using the scheme (6.4) handles the issue of being able to compare the two discrete Hessian approximations, we still have the issue of the discrete Hessian not being negative semidefinite at a relative maximum, as discussed in the following section.

6.1.2 Discrete Hessians and Relative Maxima

The fundamental difficulty in proving a FD scheme for a high-dimensional elliptic problem converges to the viscosity solution using the approach in the one-dimensional convergence proof is generalizing (2.32), or necessarily (2.33). Suppose a function v on Ω has a relative maximum at $x_0 \in \Omega$. Then, for the one-dimensional convegence proof to be generalized, we need to be able to guarantee there exists a sequence $\{h_k\}_{k\geq 0}$ such that $h_k \searrow 0$ and $\lim_{k\to\infty} \overline{D}^2_{h_k} v(x_0) \leq 0$. However, such a result is not known when v is not C^2. The inequality cannot be shown algebraically as it was in the one-dimensional proof due to the presence of the off-diagonal elements in the discrete Hessian approximation. Thus, a major hurdle that remains to be overcome for proving convergence of the high-dimensional FD method (2.68) while using the one-dimensional proof as a template is showing that the off-diagonal terms are handled correctly in the sense that the discrete Hessian approximation is negative semidefinite in limit at a relative maximum.

We now explore the behavior of the discrete Hessian approximation $\overline{D}^2_h v$ at a relative maximum. For transparency, we assume $d = 2$. Suppose $v : \Omega \to \mathbb{R}$ has a strict relative maximum at $(x_0, y_0) \in \Omega$. Then, for all sequences $\{h_{x_k}\}_{k\geq 1}$, $\{h_{y_k}\}_{k\geq 1}$ such that $h_{x_k}, h_{y_k} \to 0$ and $\lim_{k\to\infty} \overline{D}^2_{h_k} v(x_0, y_0)$ exists in $\overline{\mathbb{R}}^{2\times 2}$ for $h_k := (h_{x_k}, h_{y_k})$,

we derive sufficient conditions such that

$$\lim_{k\to\infty} \overline{D}^2_{h_k} v(x_0, y_0) \le 0,$$

where $A \le 0$ for $A = A^T \in \overline{\mathbb{R}}^{2\times 2}$ implies $-A$ is nonnegative definite.

Fix $h = (h_x, h_y)$. Observe, by (2.3),

$$\begin{aligned}
\left[\overline{D}^2_h v(x_0, y_0)\right]_{1,2} &= \left[\overline{D}^2_h v(x_0, y_0)\right]_{2,1} \\
&= \frac{v(x_0 + h_x, y_0) - v(x_0, y_0) - v(x_0 + h_x, y_0 - h_y) + v(x_0, y_0 - h_y)}{2h_x h_y} \\
&\quad + \frac{v(x_0, y_0 + h_y) - v(x_0 - h_x, y_0 + h_y) - v(x_0, y_0) + v(x_0 - h_x, y_0)}{2h_x h_y} \\
&= \frac{h_x}{2h_y}\left[\overline{D}^2_h v(x_0, y_0)\right]_{1,1} + \frac{h_y}{2h_x}\left[\overline{D}^2_h v(x_0, y_0)\right]_{2,2} \\
&\quad - \frac{v(x_0 - h_x, y_0 + h_y) - 2v(x_0, y_0) + v(x_0 + h_x, y_0 - h_y)}{2h_x h_y}.
\end{aligned}$$

Define $\delta^2_{\xi,h}$ by

$$\delta^2_{\xi,h} v(x_0, y_0) := \frac{v(x_0 - h_x, y_0 + h_y) - 2v(x_0, y_0) + v(x_0 + h_x, y_0 - h_y)}{h_x^2 + h_y^2}. \tag{6.5}$$

Then, we have

$$\left[\overline{D}^2_h v(x_0, y_0)\right]_{1,2} = \frac{h_x}{2h_y}\delta^2_{x,h_x} v(x_0, y_0) + \frac{h_y}{2h_x}\delta^2_{y,h_y} v(x_0, y_0) - \frac{h_x^2 + h_y^2}{2h_x h_y}\delta^2_{\xi,h} v(x_0, y_0).$$

By symmetry, we know both eigenvalues of $\overline{D}^2_h v(x_0, y_0)$ are real. Thus, we have $\overline{D}^2_h v(x_0, y_0) \le 0$ if and only if the two eigenvalues of $\overline{D}^2_h v(x_0, y_0)$ are nonpositive.

Let λ_h be an eigenvalue of $\overline{D}^2_h v(x_0, y_0)$ and let

$$\overline{D}^2_h v(x_0, y_0) = \begin{bmatrix} a_h & c_h \\ c_h & b_h \end{bmatrix}$$

for

$$a_h = \delta^2_{x,h_x} v(x_0, y_0), \tag{6.6a}$$

$$b_h = \delta^2_{y,h_y} v(x_0, y_0), \tag{6.6b}$$

$$c_h = \frac{h_x}{2h_y} a_h + \frac{h_y}{2h_x} b_h - \frac{h_x^2 + h_y^2}{2h_x h_y} \delta^2_{\xi,h} v(x_0, y_0). \tag{6.6c}$$

Then, λ_h satisfies

$$0 = (a_h - \lambda_h)(b_h - \lambda_h) - c_h^2 = \lambda_h^2 - (a_h + b_h)\lambda_h + a_h b_h - c_h^2.$$

Thus, we have

$$2\lambda_h = a_h + b_h \pm \sqrt{(a_h + b_h)^2 - 4(a_h b_h - c_h^2)}. \tag{6.7}$$

Using the fact that $v(x_0, y_0)$ is a strict relative maximum, there exists $\rho > 0$ such that

$$v(x, y) < v(x_0, y_0)$$

for all $(x, y) \in B_\rho(x_0, y_0)$, the ball of radius ρ centered at (x_0, y_0). Therefore,

$$\delta^2_{x,h_x} v(x_0, y_0) < 0, \qquad \delta^2_{y,h_y} v(x_0, y_0) < 0, \qquad \delta^2_{\xi,h} v(x_0, y_0) < 0, \tag{6.8}$$

for all h such that $|h| < \rho$. By (6.7), we have $\lambda_h \leq 0$ if and only if $a_h b_h \geq c_h^2$. Thus, we require $|c_h| \leq \sqrt{a_h b_h}$ since $a_h b_h > 0$. By (6.8) and (6.6), we have $|c_h| \leq \sqrt{a_h b_h}$ if and only if

$$\frac{h_x}{2h_y} a_h + \frac{h_y}{2h_x} b_h - \sqrt{a_h b_h} \leq \frac{h_x^2 + h_y^2}{2h_x h_y} \delta^2_{\xi,h} v(x_0, y_0) \leq \frac{h_x}{2h_y} a_h + \frac{h_y}{2h_x} b_h + \sqrt{a_h b_h}.$$

Thus, we require

$$\begin{aligned}
&\frac{h_x}{2h_y}\delta^2_{x,h_x}v(x_0,y_0)+\frac{h_y}{2h_x}\delta^2_{y,h_y}v(x_0,y_0)-\sqrt{\delta^2_{x,h_x}v(x_0,y_0)\delta^2_{y,h_y}v(x_0,y_0)} \qquad (6.9)\\
&\quad\le\frac{h_x^2+h_y^2}{2h_xh_y}\delta^2_{\xi,h}v(x_0,y_0)\\
&\qquad\le\frac{h_x}{2h_y}\delta^2_{x,h_x}v(x_0,y_0)+\frac{h_y}{2h_x}\delta^2_{y,h_y}v(x_0,y_0)+\sqrt{\delta^2_{x,h_x}v(x_0,y_0)\delta^2_{y,h_y}v(x_0,y_0)}
\end{aligned}$$

for $\delta^2_{\xi,h}v(x_0,y_0)$ defined by (6.5). Therefore, we have

$$\lim_{k\to\infty}\overline{D}^2_{h_k}v(x_0,y_0)\le 0$$

if and only if (6.9) holds for h_k in the limit as $k\to\infty$. Since $\overline{D}^2_{h_k}v(x_0,y_0)$ is defined in $\overline{\mathbb{R}}^{2\times 2}$, we use the convention

$$\lim_{k\to\infty}k\begin{bmatrix}-1&-1\\-1&-1\end{bmatrix}\le 0$$

since the eigenvalues are $0,-2k$ for all k.

Rescaling (6.9), we have $\lambda_h\le 0$ if and only if

$$\begin{aligned}
&\big(v(x_0-h_x,y_0)-2v(x_0,y_0)+v(x_0+h_x,y_0)\big) \qquad (6.10)\\
&\qquad+\big(v(x_0,y_0-h_y)-2v(x_0,y_0)+v(x_0,y_0+h_y)\big)-2R_h\\
&\quad\le v(x_0-h_x,y_0+h_y)-2v(x_0,y_0)+v(x_0+h_x,y_0-h_y)\\
&\qquad\le\big(v(x_0-h_x,y_0)-2v(x_0,y_0)+v(x_0+h_x,y_0)\big)\\
&\qquad\qquad+\big(v(x_0,y_0-h_y)-2v(x_0,y_0)+v(x_0,y_0+h_y)\big)+2R_h
\end{aligned}$$

for

$$\begin{aligned}
R_h^2=&\big(v(x_0-h_x,y_0)-2v(x_0,y_0)+v(x_0+h_x,y_0)\big)\\
&\times\big(v(x_0,y_0-h_y)-2v(x_0,y_0)+v(x_0,y_0+h_y)\big).
\end{aligned}$$

Therefore, we have

$$\lim_{k\to\infty} \overline{D}^2_{h_k} v(x_0, y_0) \leq 0$$

if and only if (6.10) holds for all h_k with $k >> 1$ or $\lambda_{h_k} \to 0$ with rate $O\left(\min\{h^2_{x_k}, h^2_{y_k}\}\right)$.

Remark 6.1.

(a) Observe, all three terms in (6.10) converge to zero for v continuous. However, (6.10) may not hold in limit when the scaling is reversed for v continuous. In the convergence proof, we would only have $\overline{u} - \varphi$ upper semi-continuous. We will make this observation more rigorous in the following two lemmas.

(b) We can see that (6.9) holds for $v \in C^2(\Omega)$ by using a Taylor's expansion and noting that (6.9) is equivalent to requiring $v_{xy}(x_0, y_0)^2 \leq v_{xx}(x_0, y_0)v_{yy}(x_0, y_0)$, which holds in limit for the second order Taylor's expansion of v. Thus, the discrete Hessian operator preserves the symmetry and semidefiniteness of the continuous Hessian operator D^2 when acting on v at (x_0, y_0), and it follows that the high-dimensional FD methods (2.68) and (6.4) converge when $u_h \to v \in C^2(\Omega) \cap C^0(\overline{\Omega})$, for u_h defined by (2.70), and the unique viscosity solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$.

Lemma 6.2. *Let $v : \Omega \to \mathbb{R}$ have a relative maximum at $(x_0, y_0) \in \Omega \subset \mathbb{R}^2$. Then the upper semi-continuity of v is not sufficient to guarantee there exists sequences $\{h_{x_k}\}_{k\geq 1}$, $\{h_{y_k}\}_{k\geq 1}$ such that $h_{x_k}, h_{y_k} \to 0$ and $\lim_{k\to\infty} \overline{D}^2_{h_k} v(x_0, y_0)$ exists in $\overline{\mathbb{R}}^{2\times 2}$ with*

$$\lim_{k\to\infty} \overline{D}^2_{h_k} v(x_0, y_0) \leq 0.$$

Proof. Let $(x_0, y_0) = (0, 0)$ and $\Omega = B_1(0)$, the unit ball centered at the origin. Define $v : \Omega \to \mathbb{R}$ by

$$v(x) = \begin{cases} -x^2 - y^2, & \text{if } xy = 0, \\ -x^2 - y^2 - 1, & \text{otherwise.} \end{cases}$$

Then v is upper semi-continuous on Ω. Observe,

$$\overline{D}_h^2 v(0,0) = \begin{bmatrix} -2 & -\frac{h_x}{h_y} - \frac{h_y}{h_x} + 2\frac{h_x^2+h_y^2}{h_x h_y} + \frac{1}{h_x h_y} \\ -\frac{h_x}{h_y} - \frac{h_y}{h_x} + 2\frac{h_x^2+h_y^2}{h_x h_y} + \frac{1}{h_x h_y} & -2 \end{bmatrix}.$$

Thus, for $0 < h_x, h_y << 1$, we have $\overline{D}_h^2 v(0,0)$ is not negative semidefinite, and it follows that there are no sequences $\{h_{x_k}\}_{k\geq 1}$, $\{h_{y_k}\}_{k\geq 1}$ such that $h_{x_k}, h_{y_k} \to 0$ and $\lim_{k\to\infty} \overline{D}_{h_k}^2 v(x_0, y_0) \leq 0$. The proof is complete. □

Lemma 6.3. *Let $v : \Omega \to \mathbb{R}$ have a relative maximum at $(x_0, y_0) \in \Omega \subset \mathbb{R}^2$. Then the continuity of v is not sufficient to guarantee for all sequences $\{h_{x_k}\}_{k\geq 1}$, $\{h_{y_k}\}_{k\geq 1}$ such that $h_{x_k}, h_{y_k} \to 0$ and $\lim_{k\to\infty} \overline{D}_{h_k}^2 v(x_0, y_0)$ exists in $\overline{\mathbb{R}}^{2\times 2}$, there holds*

$$\lim_{k\to\infty} \overline{D}_{h_k}^2 v(x_0, y_0) \leq 0.$$

Proof. Let $(x_0, y_0) = (0, 0)$ and $\Omega = B_1(0)$, the unit ball centered at the origin. Let $v : \Omega \to \mathbb{R}$ be continuous with

$$v(x) = \begin{cases} -x^2 - y^2, & \text{if } xy = 0, \\ -|x - y|, & \text{if } x = -y. \end{cases}$$

Suppose $h_x = h_y$. Observe,

$$\overline{D}_h^2 v(0,0) = \begin{bmatrix} -2 & -2 + \frac{2}{h_x} \\ -2 + \frac{2}{h_x} & -2 \end{bmatrix}.$$

Thus, for $0 < h_x << 1$, we have $\overline{D}_h^2 v(0,0)$ is not negative semidefinite, and it follows that $\lim_{k\to\infty} \overline{D}_{h_k}^2 v(x_0, y_0) \leq 0$ does not hold for $\{h_{x_k}\}_{k\geq 1}$ such that $h_{x_k} \to 0$ and $h_{y_k} = h_{x_k}$. The proof is complete. □

Remark 6.2.

(a) *Suppose v is an upper semi-continuous function that attains a relative maximum at (x_0, y_0). Let $\lambda_1, \lambda_2 \in \overline{\mathbb{R}}$ be the eigenvalues of $\overline{D}_h^2 v(x_0, y_0)$. Due to the fact that $\delta_{x_i,h_i}^2 v(x_0, y_0) \leq 0$, $i = 1, 2$, by (6.7), we have only two possible cases: either $\lambda_1, \lambda_2 \leq 0$ or $\lambda_i > 0$ for some $i = 1, 2$ and $\lambda_1 \lambda_2 \leq 0$. Thus, we never have $\overline{D}_h^2 v(x_0, y_0) > 0$, and it follows that the discrete Hessian operator $\overline{D}_h^2$ does not have the resolution to distinguish between a relative maximum and a saddle point when acting on low-regularity functions.*

(b) *Suppose $\overline{u} - \varphi$ attains a relative maximum at (x_0, y_0) for $\overline{u}$ an upper semi-continuous function and $\varphi \in C^2$. Since we are unable to construct a sequence such that*

$$\lim_{k\to\infty} \overline{D}_{h_k}^2 \overline{u}(x_0, y_0) \leq D^2\varphi(x_0, y_0),$$

we do not have

$$\begin{aligned} F_*\bigg(\lim_{k\to\infty} \overline{D}_{h_k}^2 \overline{u}(x_0, y_0), \nabla\varphi(x_0, y_0), \varphi(x_0, y_0), x_0, y_0 \bigg) \\ \geq F_*\bigg(D^2\varphi(x_0, y_0), \nabla\varphi(x_0, y_0), \varphi(x_0, y_0), x_0, y_0 \bigg). \end{aligned}$$

Thus, another direction would be to show

$$\lim_{k\to\infty} \Big[D_{1_{h_k}}^2 \overline{u}(x_0, y_0) \Big]_{i,j} \geq \big[D^2\varphi(x_0, y_0) \big]_{i,j} \geq \lim_{k\to\infty} \Big[D_{2_{h_k}}^2 \overline{u}(x_0, y_0) \Big]_{i,j}$$

for two discrete Hessian operators $D_{1_h}^2$ and $D_{2_h}^2$ and $i, j = 1, 2$ such that $i \neq j$. Then using the g-monotonicity assumption with component-wise ordering and consistency, we would arrive at

$$F_*\bigg(D', \nabla\varphi(x_0, y_0), \varphi(x_0, y_0), x_0, y_0 \bigg)$$

for

$$[D']_{i,j} = \begin{cases} \varphi_{x_i x_j}(x_0, y_0), & \text{for } i \neq j, \\ \liminf_{h \to 0} \delta^2_{x_i, h_i} \overline{u}(x_0, y_0), & \text{for } i = j \end{cases}$$

and $D' \leq D^2\varphi(x_0, y_0)$. *However, by Section 6.1.1, we see that such an inequality is not known.*

(c) *Similar remarks and counterexamples exist regarding the discrete Hessian operators* D_h^2, $\widehat{D}_h^2$, *and* $\widetilde{D}_h^2$ *(defined in Section 2.1) acting on a function at a relative maximum.*

6.2 The DGFE Differential Calculus and Applications

A major hurdle in approximating viscosity solutions of fully nonlinear PDEs is the necessity to approximate and/or numerically interpret weak derivatives and, in some cases, distributional derivates. To this end, a DGFE differential calculus was developed in [26]. The discrete calculus uses DG methodologies to systematically develop a numerical differentiation theory targeted towards approximating weak derivatives of Sobolev functions and piecewise Sobolev functions. In the paper, various numerical derivative operators are introduced such as the gradient, divergence, Hessian, and Laplacian operator. Then, corresponding calculus rules are established such as the product and chain rules, integration by parts formulas, and the divergence theorem. Furthermore, approximation properties are established as well as relationships with known FD numerical differential calculus theory. Using the DGFE differential calculus, classical DG methods for numerical PDEs are reinterpreted and new DG methods such as the DWDG method are developed. Additionally, by choosing V^h correctly, classical FEM can be expressed. Thus, the DGFE differential

calculus provides a unique formalism that has promising potential in the study of numerical PDEs.

6.2.1 Formulation

We now give a brief overview of the derivation and some results found in [26]. The following uses the notation defined in Section 3.1. For an interior face/edge $e \in \mathcal{E}_h^I$, we define the trace operators

$$\mathcal{Q}_i^{\pm}(v) := \{v\} \pm \frac{1}{2}\text{sgn}(n_e^{(i)})[v], \quad \text{where} \quad \text{sgn}(n_e^{(i)}) = \begin{cases} 1 & \text{if } n_e^{(i)} \geq 0, \\ -1 & \text{if } n_e^{(i)} < 0 \end{cases} \tag{6.11}$$

for all $v \in \mathcal{V}^h := W^{1,1}(\mathcal{T}_h) \cap C^0(\mathcal{T}_h)$, where $n_e = \left(n_e^{(1)}, n_e^{(2)}, \ldots, n_e^{(d)}\right)^t$ denotes the unit normal for e. Observe, the trace operators are equivalent to the trace operators $T_i^{\pm}$ defined locally in Section 3.2.2. For a boundary face/edge $e \in \mathcal{E}_h^B$, we simply let

$$\mathcal{Q}^{\pm}(v)(x) := \lim_{\substack{y \in \Omega \\ y \to x}} v(y) \qquad \forall\, x \in e. \tag{6.12}$$

Let $\gamma_i^{\pm}$ be piecewise constants with respect to the set of interior faces/edges $\mathcal{E}_h^I$. Then, using the trace operators $\mathcal{Q}^{\pm}$, we define the discrete partial derivatives $\partial_{h,x_i}^{\pm} v \in V^h$ by

$$\left(\partial_{h,x_i}^{\pm} v, \varphi_h\right)_{\mathcal{T}_h} := \left\langle \mathcal{Q}_i^{\pm}(v)\, n^{(i)}, [\varphi_h] \right\rangle_{\mathcal{E}_h} - \left(v, \partial_{x_i}\varphi_h\right)_{\mathcal{T}_h} + \left\langle \gamma_i^{\pm}[v], [\varphi_h] \right\rangle_{\mathcal{E}_h^I} \quad \forall \varphi_h \in V^h, \tag{6.13}$$

and, if v has known boundary data $g \in L^1(\partial\Omega)$, we define the discrete partial derivatives $\partial_{h,x_i}^{\pm,g} v \in V^h$ by

$$\left(\partial_{h,x_i}^{\pm,g} v, \varphi_h\right)_{\mathcal{T}_h} := \left(\partial_{h,x_i}^{\pm} v, \varphi_h\right)_{\mathcal{T}_h} + \left\langle g\, n^{(i)}, \varphi_h \right\rangle_{\widetilde{\mathcal{E}}_h^B} - \left\langle v\, n^{(i)}, \varphi_h \right\rangle_{\mathcal{E}_h^B} \qquad \forall \varphi_h \in V^h. \tag{6.14}$$

Using the above discrete partial derivate operators, we can immediately define first and second order discrete derivate operators. To this end, we define the first order discrete derivative operators by

$$\nabla^{\pm}_{h,*} v := \left(\partial^{\pm,*}_{h,x_1} v, \partial^{\pm,*}_{h,x_2} v, \cdots, \partial^{\pm,*}_{h,x_d} v\right)^t, \tag{6.15}$$

$$\mathrm{div}^{\pm}_{h,*} \vec{v} := \partial^{\pm,*}_{h,x_1} v_1 + \partial^{\pm,*}_{h,x_2} v_2 + \cdots + \partial^{\pm,*}_{h,x_d} v_d \tag{6.16}$$

for any $v \in \mathcal{V}^h$ and $\vec{v} = (v_1, v_2, \cdots, v_d)^t \in \left[\mathcal{V}^h\right]^d$. We also define the second order discrete derivative operators by

$$D^{+,\pm}_{h;*,**} v := \nabla^{+}_{h,**} \left(\nabla^{\pm}_{h,*} v\right)^t, \qquad D^{-,\pm}_{h;*,**} v := \nabla^{-}_{h,**} \left(\nabla^{\pm}_{h,*} v\right)^t, \tag{6.17}$$

$$\Delta^{+,\pm}_{h;*,**} v := \mathrm{div}^{+}_{h,**} \nabla^{\pm}_{h,*} v, \qquad \Delta^{-,\pm}_{h;*,**} v := \mathrm{div}^{-}_{h,**} \nabla^{\pm}_{h,*} v, \tag{6.18}$$

for any $v \in \mathcal{V}^h$. In the above definitions, $*, **$ can either be empty or denote a known boundary data function, where $*$ corresponds to Dirichlet data and $**$ corresponds to Neumman data.

6.2.2 Properties of DGFE Discrete Derivatives

We now state some properties concerning the DGFE discrete derivatives defined in Section 6.2.1.

Proposition 6.1. *For any $v \in \mathcal{V}^h \cap H^1(\Omega)$, $\partial^{\pm}_{h,x_i} v$ coincides with the L^2-projection of $\partial_{x_i} v$ onto V^h. We write $\partial^{\pm}_{h,x_i} v = \mathcal{P}_h \partial_{x_i} v$, where $\mathcal{P}_h$ denotes the L^2 projection onto V^h.*

Theorem 6.1. *Let $F \in C^1(\mathbb{R}), F' \in L^\infty(\mathbb{R})$. For $u, v \in \mathcal{V}^h \cap C^0(\Omega)$, there holds, for $i = 1, 2, \cdots, d$,*

$$\partial_{h,x_i}(uv) = \mathcal{P}_h\big(u \partial_{x_i} v + v \partial_{x_i} u\big), \tag{6.19}$$

$$\partial_{h,x_i} F(u) = \mathcal{P}_h\big(F'(u) \partial_{x_i} u\big), \tag{6.20}$$

where

$$\partial_{h,x_i} := \frac{\partial^+_{h,x_i} + \partial^-_{h,x_i}}{2}.$$

Theorem 6.2. *The formal adjoint of the operator* $\mathrm{div}_h^{\pm}$ *is* $-\nabla^{\mp}_{h,0}$ *with respect to the inner product* $(\cdot,\cdot)_{\mathcal{T}_h}$ *provided* $\gamma_i^+ = -\gamma_i^-$ *for all* $i = 1, 2, \ldots, d$*; that is,*

$$(\mathrm{div}_h^{\pm}\vec{v}_h, \varphi_h)_{\mathcal{T}_h} = -(\vec{v}_h, \nabla^{\mp}_{h,0}\varphi_h)_{\mathcal{T}_h} \tag{6.21}$$

for all $\vec{v}_h \in \left[\mathcal{V}^h\right]^d$, $\varphi_h \in V^h$*. In addition, if* $\gamma_i^+ = -\gamma_i^-$*, then the formal adjoint of the operator* $\mathrm{div}^{\pm}_{h,\vec{0}}$ *is* $-\nabla^{\mp}_h$*.*

Remark 6.3. *The above properties can all be extended for functions from the space* $\mathcal{V}^h$*. See [26] for more details.*

6.2.3 The DWDG Method

We end this section by formulating a new DG method that was first inspired by the observations in Section 3.2.5 and later analyzed in the context of Poisson's equation using the formalism of the DGFE discrete calculus. The following material can be found in [40].

Consider the Poisson equation with Dirichlet boundary condition:

$$-\Delta u = f \qquad \text{in } \Omega, \tag{6.22a}$$

$$u = g \qquad \text{on } \partial\Omega, \tag{6.22b}$$

where $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ are two given functions. Let $j_{h,g} : \mathcal{V}^h \to V^h$ be the unique operator satisfying

$$\big(j_{h,g}(v), \varphi_h\big)_{\mathcal{T}_h} = \big\langle \eta_1[v], [\varphi_h]\big\rangle_{\mathcal{E}_h^I} + \big\langle \eta_1 v, \varphi_h\big\rangle_{\mathcal{E}_h^B} - \big\langle \eta_1 g, \varphi_h\big\rangle_{\widetilde{\mathcal{E}}_h^B} \qquad \forall \varphi_h \in V^h, \tag{6.23}$$

where η_1 is a penalty parameter that is piecewise constant with respect to the set of faces/edges, and suppose $\gamma_i^{\pm} = 0$ in (6.13). Then, the symmetric DWDG method is

given by

$$-\frac{\Delta_{h,g}^{-+}u_h + \Delta_{h,g}^{+-}u_h}{2} + j_{h,g}(u_h) = \mathcal{P}_h f. \tag{6.24}$$

By Theorem 6.2, problem (6.24) is equivalent to finding $u_h \in V^h$ such that

$$\frac{1}{2}(\nabla_{h,g}^{+}u_h, \nabla_{h,0}^{+}v_h)_{\mathcal{T}_h} + \frac{1}{2}(\nabla_{h,g}^{-}u_h, \nabla_{h,0}^{-}v_h)_{\mathcal{T}_h} + j_{h,g}(u_h) = (f, v_h)_{\mathcal{T}_h} \tag{6.25}$$

for all $v_h \in V^h$.

Theorem 6.3. *Set $\eta_{\min} := \min_{e\in\mathcal{E}_h} h_e^{-1}\eta_1(e)$. Suppose that there exists at least one simplex $K \in \mathcal{T}_h$ with exactly one boundary face/edge. Then, there exists a unique solution to (6.24) provided $\eta_{\min} \geq 0$. Furthermore, if the triangulation is quasi-uniform, and if each simplex in the triangulation has at most one boundary face/edge, then there exists a constant $C_* > 0$ independent of h and η_1 such that problem (6.24) has a unique solution provided $\eta_{\min} > -C_*$.*

Theorem 6.4. *Let u_h be the solution to (6.24), $u \in H^{s+1}(\Omega)$ be the solution to (6.22), and $\eta_{\max} = \max_{e\in\mathcal{E}_h} h_e^{-1}\eta_1(e)$. Then u_h satisfies the following estimate provided $\eta_{\min} > 0$:*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{s+1}\Big(\sqrt{\eta_{\max}} + \frac{1}{\sqrt{\eta_{\min}}}\Big)^2 |u|_{H^{s+1}(\Omega)} \qquad (1 \leq s \leq r), \tag{6.26}$$

and if the triangulation is quasi-uniform and $\eta_{\min} > -C_*$, then there holds

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{s+1}\Big(\sqrt{|\eta_{\min}|} + \frac{1}{\sqrt{C_* + \eta_{\min}}}\Big)^2 |u|_{H^{s+1}(\Omega)}, \tag{6.27}$$

where C denotes a generic positive constant independent of h, and C_* is the positive constant from Theorem 6.3.

We emphasize that problem (6.24) is well-posed without the added penalty term $j_{h,g}$. As far as we are aware, this is the first symmetric DG method that has this

property in any dimension (cf. [40] and the references therein). The optimality of the DWDG method for the Poisson equation and the lack of a penalty term warrants applying the DWDG method to other second order PDEs, especially since the lack of a penalty term may have positive implications when using multigrid solvers. The formulation can also easily be extended to the fourth order biharmonic equation.

6.3 Linear Second Order Elliptic Equations of Non-Divergence Form

In order to apply Algorithm 5.1 for approximating the Hamilton-Jacobi-Bellman equation, we must first be able to approximate the viscosity solutions of linear second order elliptic equations of non-divergence form. Thus, we need to develop convergent numerical schemes for the problem

$$Lu := -A : D^2 u + \vec{b} \cdot \nabla u + c\, u = f \qquad \text{in } \Omega, \tag{6.28a}$$

$$u = g \qquad \text{on } \partial\Omega, \tag{6.28b}$$

where we assume $A \in [L^\infty(\Omega)]^{d\times d}$; $\vec{b} \in [L^\infty(\Omega)]^d$; $c, f \in L^\infty(\Omega)$, $c \geq 0$; and there exists $\lambda, \Lambda > 0$ such that

$$\lambda\, |\xi|^2 \leq \xi \cdot A(x)\xi \leq \Lambda\, |\xi|^2 \qquad \forall\, \xi \in \mathbb{R}^d, x \in \Omega.$$

As far as we are aware, no existing numerical methods are known to converge to the viscosity solution of (6.28) without making additional assumptions (cf. [50] and the references therein).

The numerical methods developed in Chapters 2, 3, and 4 are designed to directly approximate the viscosity solution of (6.28). Furthermore, when A is a diagonal matrix, we have shown the FD methods in Chapter 2 are convergent. Also, by the L^∞ constraints on the differential operator L, the Lax-Friedrichs-like numerical operators

are globally g-monotone. However, the admissibility and stability proofs in Chapter 2 for the Lax-Friedrichs-like numerical operators would need to be reworked to account for the low order terms. In particular, the fixed point argument used is based on forming a mapping that commutes with the addition of constants, a property that no longer holds if $c \neq 0$. We do note that the form of the Hamilton-Jacobi-Bellman equation that arises from stochastic optimal control, as formulated in Section 1.4.2, does in fact have $c = 0$.

An indirect approach for approximating (6.28) was developed in Chapter 5 using the vanishing moment method. However, the analysis assumed A is continuous, $\vec{b} = \vec{0}$, and $c = 0$, in which case the viscosity solution is actually in $H^2(\Omega)$. Thus, the stability and convergence analysis for the vanishing moment method needs to be extended to account for lower order terms and a lack of continuity for the coefficient functions. To this end, we would expect the stability estimates for u^ϵ to be inversely related to ϵ in both the H^1 and H^2 semi-norms. However, the stability estimates for u^ϵ may still be independent of ϵ in the L^2 norm, allowing for the weak passage of limits in the existence proof and possibly the same rates of convergence for u^ϵ to the viscosity solution u in the L^2 norm with respect to ϵ. Since the underlying viscosity solution would be in C^0, estimates should also be derived in the C^0 or L^∞ norms.

The analysis for the vanishing moment method also required the use of a couple of conjectures that need to be proved. The two conjectures both resemble the trace inequality with additional jump terms for the piecewise constant matrix-valued function $\overline{A}$. We now consider proving Conjecture 5.1 using the trace inequality from

Section 5.2.1 with $\alpha = 0$:

$$
\begin{aligned}
\sum_{j=1}^{M} (A_j \nabla v \cdot n_j, v)_{\partial\Omega_j} &= \sum_{\Gamma_j \in \mathcal{E}^B} (A_j \nabla v \cdot n_j, v)_{\Gamma_j} \\
&\quad + \sum_{\Gamma_{i,j} \in \mathcal{E}^I} \Big((A_i - A_j) \nabla v \cdot n_i, v \Big)_{\Gamma_{i,j}} \\
&= \sum_{\Gamma_{i,j} \in \mathcal{E}^I} \Big((A_i - A_j) \nabla v \cdot n_i, v \Big)_{\Gamma_{i,j}} \\
&\leq \sum_{\Gamma_{i,j} \in \mathcal{E}^I} \left| \Big((A_i - A_j) \nabla v \cdot n_i, v \Big)_{\Gamma_{i,j}} \right| \\
&\leq \sum_{\Gamma_{i,j} \in \mathcal{E}^I} \| (A_i - A_j) \nabla v \cdot n_i \|_{L^2(\Gamma_{i,j})} \, \| v \|_{L^2(\Gamma_{i,j})} \\
&\leq \sum_{\Gamma_{i,j} \in \mathcal{E}^I} \| (A_i - A_j) \|_{L^\infty(\Gamma_{i,j})} \, \| \nabla v \cdot n_i \|_{L^2(\Gamma_{i,j})} \, \| v \|_{L^2(\Gamma_{i,j})} \\
&\leq \delta_A \sum_{\Gamma_{i,j} \in \mathcal{E}^I} \| \nabla v \|_{L^2(\Gamma_{i,j})} \, \| v \|_{L^2(\Gamma_{i,j})} \\
&\leq \delta_A \sum_{j=1}^{M} \| \nabla v \|_{L^2(\partial\Omega_j)} \, \| v \|_{L^2(\partial\Omega_j)} \\
&\leq \frac{\delta_A}{2} \sum_{j=1}^{M} \| \nabla v \|^2_{L^2(\partial\Omega_j)} + \frac{\delta_A}{2} \sum_{j=1}^{M} \| v \|^2_{L^2(\partial\Omega_j)} \\
&\leq \frac{\delta_A C_T^2}{2} \sum_{j=1}^{M} \| \nabla v \|^2_{H^1(\Omega_j)} + \frac{\delta_A C_T^2}{2} \sum_{j=1}^{M} \| v \|^2_{H^1(\Omega_j)} \\
&\leq \frac{\delta_A C_T^2}{2} \left\| D^2 v \right\|^2_{L^2(\Omega)} + \delta_A C_T^2 \| \nabla v \|^2_{L^2(\Omega)} \\
&\quad + \frac{\delta_A C_T^2}{2} \| v \|^2_{L^2(\Omega)} \\
&\leq \delta_A C_T^2 \| v \|^2_{H^2(\Omega)}
\end{aligned}
\tag{6.29}
$$

Unfortunately, the inequality is not sufficiently sharp due to the fact the constant C_T is inversely related to $\rho \approx \operatorname{diam} \Omega_j$ for all $j = 1, 2, \ldots, M$. Thus, we cannot use δ_A to control the right-hand side of (6.29), and it follows that we cannot use the above

estimate in the a-priori estimates of Chapter 5 without a stronger assumption on the regularity of A such as a Lipschitz continuity assumption.

We end with a couple more comments regarding sharpening (6.29). First, the above estimate does not need to hold for all $v \in H^2(\Omega) \cap H_0^1(\Omega)$, but instead just the eigenfunctions of the Laplace operator. The constant from Poincarè's inequality is directly related to ρ, and $\Delta^j \psi$ has zero trace for all integers $j \geq 0$ whenever ψ is an eigenfunction of the Laplace operator. Second, the strong solution u for (5.8) can be shown to be uniformly bounded in H^2 using cutoff functions (cf. [32]). Thus, it may be possible to derive uniform H^2 estimates for the strong solution u using the technique where A is approximated by $\overline{A}$, and uniform estimates for u^ϵ would immediately follow. Then, the vanishing moment method would amount to a numerical technique to avoid stability issues such as those in Example 5.3 when using $\epsilon = 0$.

Bibliography

[1] Robert A. Adams and John J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003. 221

[2] A. D. Aleksandrov. Certain estimates for the Dirichlet problem. *Soviet Math. Dokl.*, 1:1151–1154, 1961. 18

[3] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Anal.*, 4(3):271–283, 1991. 4, 5, 16, 39

[4] Guy Barles and Espen R. Jakobsen. Error bounds for monotone approximation schemes for parabolic Hamilton-Jacobi-Bellman equations. *Math. Comp.*, 76(260):1861–1893, 2007. 16

[5] Klaus Böhmer. On finite element methods for fully nonlinear elliptic equations of second order. *SIAM J. Numer. Anal.*, 46(3):1212–1249, 2008. 16

[6] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008. 237

[7] Luis A. Caffarelli and Xavier Cabré. *Fully nonlinear elliptic equations*, volume 43 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 1995. 4, 5

[8] Luis A. Caffarelli and Panagiotis E. Souganidis. A rate of convergence for monotone finite difference approximations to fully nonlinear, uniformly elliptic PDEs. *Comm. Pure Appl. Math.*, 61(1):1–17, 2008. 39

[9] Philippe G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original. 92, 124, 251

[10] Bernardo Cockburn and Chi-Wang Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.*, 35(6):2440–2463, 1998. 92

[11] COMSOL Multiphysics. *version 3.5a*. Comsol Group, Stockholm, Sweden, 2008. 23

[12] COMSOL Multiphysics. *version 4.0a*. Comsol Group, Stockholm, Sweden, 2010. 23

[13] M. G. Crandall, L. C. Evans, and P.-L. Lions. Some properties of viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.*, 282(2):487–502, 1984. 5, 6

[14] M. G. Crandall and P.-L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Math. Comp.*, 43(167):1–19, 1984. 9, 12, 15, 33

[15] Michael G. Crandall and Hitoshi Ishii. The maximum principle for semicontinuous functions. *Differential Integral Equations*, 3(6):1001–1014, 1990. 10

[16] Michael G. Crandall, Hitoshi Ishii, and Pierre-Louis Lions. User's guide to viscosity solutions of second order partial differential equations. *Bull. Amer. Math. Soc. (N.S.)*, 27(1):1–67, 1992. 10

[17] Michael G. Crandall and Pierre-Louis Lions. Viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.*, 277(1):1–42, 1983. 5, 9

[18] Michael G. Crandall and Luc Tartar. Some relations between nonexpansive and order preserving mappings. *Proc. Amer. Math. Soc.*, 78(3):385–390, 1980. 54

[19] Edward J. Dean and Roland Glowinski. On the numerical solution of a two-dimensional Pucci's equation with Dirichlet boundary conditions: a least-squares approach. *C. R. Math. Acad. Sci. Paris*, 341(6):375–380, 2005. 14

[20] Edward J. Dean and Roland Glowinski. On the numerical solution of the elliptic Monge-Ampère equation in dimension two: a least-squares approach. In *Partial differential equations*, volume 16 of *Comput. Methods Appl. Sci.*, pages 43–63. Springer, Dordrecht, 2008. 14

[21] Lawrence C. Evans. A convergence theorem for solutions of nonlinear second-order elliptic equations. *Indiana Univ. Math. J.*, 27(5):875–887, 1978. 10

[22] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010. 220, 224, 229

[23] X. Feng and T. Lewis. Local discontinuous galerkin methods for one-dimensional second order fully nonlinear elliptic and parabolic equations. *Preprint.* arxiv.org/abs/1212.0537, 2013. 23

[24] X. Feng and T. Lewis. Mixed interior penalty discontinuous galerkin methods for fully nonlinear second order elliptic and parabolic equations in high dimensions. *Preprint*, 2013. 23

[25] X. Feng and T. Lewis. Mixed interior penalty discontinuous galerkin methods for one-dimensional fully nonlinear second order elliptic and parabolic equations. *Preprint.* arxiv.org/abs/1212.0259, 2013. 23

[26] X. Feng, T. Lewis, and M. Neilan. Discontinuous galerkin finite element differential calculus and applications to numerical solutions of linear and nonlinear partial differential equations. *Preprint.* arxiv.org/abs/1302.6984, 2013. 23, 94, 107, 113, 122, 275, 276, 278

[27] Xiaobing Feng, Roland Glowinski, and Michael Neilan. Recent Developments in Numerical Methods for Fully Nonlinear Second Order Partial Differential Equations. *SIAM Rev.*, 55(2):205–267, 2013. 11, 15, 16, 142, 203

[28] Xiaobing Feng, Chiu-Yen Kao, and Thomas Lewis. Convergent finite difference methods for one-dimensional fully nonlinear second order partial differential equations. *J. Comput. Appl. Math.*, 254:81–98, 2013. 23

[29] Xiaobing Feng and Michael Neilan. Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. *J. Sci. Comput.*, 38(1):74–98, 2009. 13, 95, 97, 107, 132

[30] Wendell H. Fleming and Raymond W. Rishel. *Deterministic and stochastic optimal control*. Springer-Verlag, Berlin, 1975. Applications of Mathematics, No. 1. 19

[31] Wendell H. Fleming and H. Mete Soner. *Controlled Markov processes and viscosity solutions*, volume 25 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1993. 19

[32] David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition. 4, 5, 18, 216, 217, 224, 240, 283

[33] Cristian E. Gutiérrez. *The Monge-Ampère equation*. Progress in Nonlinear Differential Equations and their Applications, 44. Birkhäuser Boston Inc., Boston, MA, 2001. 18

[34] Jan S. Hesthaven and Tim Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008. Algorithms, analysis, and applications. 93

[35] Hitoshi Ishii. On uniqueness and existence of viscosity solutions of fully nonlinear second-order elliptic PDEs. *Comm. Pure Appl. Math.*, 42(1):15–45, 1989. 10

[36] Max Jensen and Iain Smears. On the Convergence of Finite Element Methods for Hamilton–Jacobi–Bellman Equations. *SIAM J. Numer. Anal.*, 51(1):137–162, 2013. 16

[37] Robert Jensen. The maximum principle for viscosity solutions of fully nonlinear second order partial differential equations. *Arch. Rational Mech. Anal.*, 101(1):1–27, 1988. 10

[38] N. Krylov. Rate of convergence of difference approximations for uniformly nondegenerate elliptic bellman's equations. *Preprint.* arxiv.org/abs/1203.2905, 2013. 16

[39] Hung Ju Kuo and Neil S. Trudinger. Discrete methods for fully nonlinear elliptic equations. *SIAM J. Numer. Anal.*, 29(1):123–135, 1992. 39

[40] T. Lewis and M. Neilan. Convergence analysis of a symmetric dual-wind discontinuous galerkin method. *Submitted*, 2013. 23, 113, 278, 280

[41] P.-L. Lions. Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. I. The dynamic programming principle and applications. *Comm. Partial Differential Equations*, 8(10):1101–1174, 1983. 10

[42] Pierre-Louis Lions. Some recent results in the optimal control of diffusion processes. In *Stochastic analysis (Katata/Kyoto, 1982)*, volume 32 of *North-Holland Math. Library*, pages 333–367. North-Holland, Amsterdam, 1984. 10

[43] Matlab. *version 7.11.1.866 (R2010b) Service Pack 1.* The MathWorks Inc., Natick, Massachusetts, 2010. 23

[44] John W. Milnor. *Topology from the differentiable viewpoint.* Based on notes by David W. Weaver. The University Press of Virginia, Charlottesville, Va., 1965. 18

[45] J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36:9–15, 1971. Collection of articles dedicated to Lothar Collatz on his sixtieth birthday. 117

[46] Adam M. Oberman. Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. *Discrete Contin. Dyn. Syst. Ser. B*, 10(1):221–238, 2008. 16

[47] A. V. Pogorelov. *Monge-Ampère equations of elliptic type*. Translated from the first Russian edition by Leo F. Boron with the assistance of Albert L. Rabenstein and Richard C. Bollinger. P. Noordhoff Ltd., Groningen, 1964. 18

[48] A. V. Pogorelov. *Extrinsic geometry of convex surfaces*. American Mathematical Society, Providence, R.I., 1973. Translated from the Russian by Israel Program for Scientific Translations, Translations of Mathematical Monographs, Vol. 35. 18

[49] Chi-Wang Shu. High order numerical methods for time dependent Hamilton-Jacobi equations. In *Mathematics and computation in imaging science and information processing*, volume 11 of *Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.*, pages 47–91. World Sci. Publ., Hackensack, NJ, 2007. 35

[50] I. Smears and E. Süli. Discontinuous galerkin finite element approximation of non-divergence form elliptic equations with cordès coefficients. *Preprint.* eprints.maths.ox.ac.uk/1623, 2012. 217, 280

[51] Eitan Tadmor. Approximate solutions of nonlinear conservation laws and related equations. In *Recent advances in partial differential equations, Venice 1996*, volume 54 of *Proc. Sympos. Appl. Math.*, pages 325–368. Amer. Math. Soc., Providence, RI, 1998. 15

[52] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. 19

[53] Jue Yan and Stanley Osher. A local discontinuous Galerkin method for directly solving Hamilton-Jacobi equations. *J. Comput. Phys.*, 230(1):232–244, 2011. 15, 92

Vita

Thomas Lewis was born in Decatur, Georgia to John and Tomi Lewis in 1985. He graduated from Parkview High School in Lilburn, Georgia in 2003. During the fall of the same year, Thomas began attending Georgia College and State University in Milledgeville, Georgia. He graduated with a Bachelor of Science degree in mathematics and minors in physics and computer science in the Spring of 2007. The following fall, he began his graduate studies in mathematics at the University of Tennessee - Knoxville. Thomas completed the requirements for the Doctorate in Philosophy degree at the University of Tennessee in August, 2013.